Large scale convex composite optimization: duality, algorithms and implementations

Defeng Sun, Kim-Chuan Toh

Department of Mathematics, National University of Singapore

Students/postdocs: Xudong Li (NUS), Ying Cui (NUS), Xin-Yee Lam (NUS)

Past contributors: Chao Ding (CAS), Xinyuan Zhao (Beijing U Tech.), Caihua Chen (Nanjing), Liang Chen (Hunan U), Yongjin Liu (Shengyang Aerospace U), Chengjing Wang (Southwest Jiaotong U), Junfeng Yang (Nanjing), Min Li (Southeast U), Liuqin Yang (Grab), Kaifeng Jiang (DBS)

Outline

- Convex functions, proximal mapping, Moreau-Yosida regularization, spectral operators and its differential properties
- Duality and examples
- An inexact symmetric Gauss-Seidel iteration
- Inexact APG (accelerated proximal gradient)
- SSNCG (semismooth Newton-CG method)
- Danskin-type Theorem
- Inexact ABCD (accelerated block coordinate descent)
- General framework of proximal-point algorithm (PPA)
- Inexact majorized semi-proximal ADMM
- SSNCG based augmented Lagrangian method (SSNAL)

Convex functions, proximal mapping, Moreau-Yosida regularization, spectral operators and its differential properties

- R. Bhatia, Matrix Analysis, Springer-Verlag, New York, 1997.
- C. Ding, An Introduction to A Class of Matrix Optimization Problems, PhD Thesis, National University of Singapore, 2012.
- F. H. Clarke, Optimization and Nonsmooth Analysis, John Wiley and Sons, New York, 1983.
- R. T. Rockafellar. *Convex Analysis*. Princeton New Jersey, 1970.
- E. H. Zarantonello. Projections on convex sets in Hilbert space and spectral theory I and II., *Contributions to Nonlinear Functional Analysis* (E. H. Zarantonello, ed.), Academic Press, New York, 1971, 237–424.

C: a subset of finite-dim real Euclidean space $\mathcal{E}$ with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$.

B_ε = ball with center at 0 and radius ε.

affine hull $\mathrm{aff}\, C = \left\{ \sum_{k=1}^{m} \lambda_k x^k \mid x^1, \ldots, x^m \in C, \sum_{k=1}^{m} \lambda_k = 1,\ m \geq 1 \right\}$

relative int. $\mathrm{ri}\, C = \{ x \in \mathrm{aff}\, C \mid \exists\, \varepsilon > 0,\ (x + B_\varepsilon) \cap \mathrm{aff}\, C \subset C \}$

If C is convex, then $\mathrm{aff}\,(\mathrm{ri}\, C) = \mathrm{aff}\, C = \mathrm{aff}\,(\mathrm{cl}(C))$, where $\mathrm{cl}(C)$ is the closure

Let $f : \mathcal{E} \to [-\infty, \infty]$ be an extended real-valued function

epigraph $\quad \mathrm{epi}\, f = \left\{ (x, \mu) \in \mathcal{E} \times \mathbb{R} \mid f(x) \leq \mu \right\}$

effective dom. $\quad \mathrm{dom}(f) = \{ x \in \mathcal{E} \mid f(x) < \infty \}$

f is said to be convex (closed) if $\mathrm{epi}\, f$ is convex (closed).
A convex function is proper if $\mathrm{epi}\, f \neq \emptyset$ and $f(x) > -\infty\ \forall\, x \in \mathcal{E}$.

Convex functions

Let C be a convex set in $\mathcal{E}$. A function $f : C \to \mathbb{R}$ is said to be

convex $\quad f(\lambda x + \beta y) \leq \lambda f(x) + \beta f(y) \; \forall \, x, y \in C, \lambda + \beta = 1, \lambda, \beta \geq 0$

Strongly convex $\exists \, \sigma > 0$ s.t. $f(\lambda x + \beta y) \leq \lambda f(x) + \beta f(y) - \frac{\sigma}{2} \lambda \beta \|x - y\|^2$

directional deriv. at x along h: $\quad f'(x; h) = \lim_{t \downarrow 0} \frac{f(x+th) - f(x)}{t}$

If f is convex and $f(x)$ is finite, then $f'(x; h)$ exists for any $h \in \mathcal{E}$.

Definition 1

A vector $x^* \in \mathcal{E}$ is said to be a subgradient of f at $x \in \mathcal{E}$ if

$$f(z) \; \geq \; f(x) + \langle x^*, \, z - x \rangle \;\; \forall \, z \in \mathcal{E}.$$

subdifferential of f at x, $\partial f(x) =$ set of all subgradients of f at x.

Let f be an extended real-valued convex fun. on $\mathcal{E}$. Then:

Proposition 2 (Rockafellar70, Sec 23–25,27,31)

1. *f proper $\Rightarrow$ ri(dom(f)) $\neq \emptyset$ and $\partial f(x) \neq \emptyset$ for any $x \in$ ri(dom(f)). Moreover, $\partial f(x)$ is bounded iff $x \in \mathrm{int}$(dom f).*
2. *f proper $\Rightarrow \partial f$ is monotone, i.e., for any $x, y \in \mathcal{E}$,*

$$\langle x - y, u - v \rangle \geq 0 \quad \forall u \in \partial f(x),\ v \in \partial f(y)$$

3. *If f is proper and convex, then f is Lipschitz cont. on any closed bounded subset of ri(dom f).*
4. *Let f, g be proper convex fun. on $\mathcal{E}$. If ri(dom f) $\cap$ ri(dom g) $\neq \emptyset$, then $\partial(f+g)(x) = \partial f(x) + \partial g(x)$ for all $x \in \mathcal{E}$.*
5. *If f is closed and proper, then $\inf_{z \in \mathcal{E}} f(z)$ is attained at x iff $0 \in \partial f(x)$.*
6. *If f is closed and proper, then ∂f is upper semicontinuous, i.e., for any $v_k \in \partial f(x_k)$ s.t. $v_k \to v$ and $x_k \to x$, we have $v \in \partial f(x)$.*
7. *If $f(x)$ is finite, then f is differentiable at x iff $\partial f(x)$ is a singleton.*

Fenchel conjugate

Let f be an extended real-valued fun. on $\mathcal{E}$.

Fenchel conj. $\quad f^*(y) = \sup\{\langle y, x\rangle - f(x) \mid x \in \mathcal{E}\}, \quad y \in \mathcal{E}.$

f^* is always closed and convex.

Proposition 3

Let f be a closed proper convex fun. on $\mathcal{E}$. For any $x \in \mathcal{E}$, we have the following equivalent conditions for a vector $x^* \in \mathcal{E}$:

1. $f(x) + f^*(x^*) = \langle x, x^*\rangle$
2. $x^* \in \partial f(x)$
3. $x \in \partial f^*(x^*)$
4. $\langle x, x^*\rangle - f(x) = \max_{z\in\mathcal{E}}\{\langle z, x^*\rangle - f(z)\}$
5. $\langle x, x^*\rangle - f^*(x^*) = \max_{z^*\in\mathcal{E}}\{\langle x, z^*\rangle - f^*(z^*)\}$

For any $C \subset \mathcal{E}$, define

indicator fun. $\quad \delta_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise} \end{cases}$

Dual cone

Let $C \subset \mathcal{E}$ be a cone, i.e. $x \in C$ and $\lambda > 0 \Rightarrow \lambda x \in C$.

$$\text{dual cone} \quad C^* = \{y \in \mathcal{E} \mid \langle y, x\rangle \geq 0 \; \forall \, x \in C\}$$

C^* is a closed convex cone (even if C is not convex).

If $C = L$, a vector subspace of $\mathcal{E}$, then $C^* = L^\perp$.

Proposition 4 (Rockafellar70, Secs 12–13)

Let f be cpc (closed proper convex) on $\mathcal{E}$. Suppose $f(0) = 0$ and f is positively homogeneous, i.e., $f(\lambda x) = \lambda f(x) \; \forall \, x \in \mathcal{E}$ and $\lambda > 0$. Then $f^* = \delta_{\partial f(0)}$.

(1) If C is also closed, then $f = \delta_C$ is positively homo. and $f(0) = 0$.

Hence $\quad \partial f(0) = C^\circ := -C^*, \quad \delta_C^* = f^* = \delta_{C^\circ}$

(2) If $f(x) = \max\{x_1, \ldots, x_n\}$, then $f(0) = 0$ and f is positively homo. with

$$\partial f(0) = \{x \in \mathbb{R}^n \mid x_1 + \ldots x_n = 1, x \geq 0\}$$

Definition 5 (Semismooth)

Let f be a locally Lipschitz cont. fun. from an open set Ω to $\mathbb{R}$.

1. *f is said to be semismooth at x if it is directionally differentiable at x and*

$$f(x+h) - f(x) - \langle \nabla f(x+h), h \rangle = o(\|h\|) \;\; \forall\, h \to 0, \; x+h \in \mathcal{D}_f$$

   where $\mathcal{D}_f = \{y \in \Omega \mid f \text{ is differentiable at } y\}$.

2. *Moreover, f is said to be strongly semismooth at x if*

$$f(x+h) - f(x) - \langle \nabla f(x+h), h \rangle = O(\|h\|^2) \;\; \forall\, h \to 0, \; x+h \in \mathcal{D}_f$$

By Rademacher's thm., f is differentiable almost everywhere on Ω.

Theorem 6

Any convex function $f : \mathcal{E} \to \mathbb{R}$ is semismooth.

Proof. By Prop. 2, f is locally Lipschitz cont. on $\mathcal{E}$. Since f is convex and $f(x)$ is finite, $f'(x;h)$ exists for all $x \in \mathcal{E}$ and $h \in \mathcal{E}$. From convexity of f, we get

$$f(x+h) - f(x) - \langle \nabla f(x+h),\, h\rangle \leq 0 \quad \forall\, x+h \in \mathcal{D}_f.$$

Since f is convex and proper, given any $x \in \mathrm{int}(\mathrm{dom}\, f)$ and $\varepsilon > 0$, $\exists \delta > 0$ such that $\partial f(x+h) \subset \partial f(x) + B_\varepsilon$ for all $\|h\| \leq \delta$. In particular when $x+h \in \mathcal{D}_f$, there exist $V_h \in \partial f(x)$ such that

$$\|\nabla f(x+h) - V_h\| \leq \varepsilon$$

$$f(x+h) - f(x) - \langle \nabla f(x+h),\, h\rangle \geq \langle V_h - \nabla f(x+h),\, h\rangle \geq -\varepsilon\|h\|$$

Hence $f(x+h) - f(x) - \langle \nabla f(x+h),\, h\rangle = o(\|h\|)$ for $x+h \in \mathcal{D}_f$, $h \to 0$.

Moreau-Yosida regularization and proximal mapping

Let $f : \mathcal{E} \to (-\infty, \infty]$ be a closed proper convex (cpc) fun.

MY reg. of f at x: $M_f(x) = \min_{y \in \mathcal{E}} \left\{ \phi(y; x) := f(y) + \frac{1}{2}\|y - x\|^2 \right\}$

Proximal mapping f at x: $P_f(x) = \text{argmin}_{y \in \mathcal{E}} \{ \cdots \}$

Proposition 7

$P_f(x)$ exists and is unique.

Proof. Uniqueness follows from strong convexity of the function $\phi(\cdot\,; x)$. Since f is proper, $\exists\ y_0 \in \text{ri}\,(\text{dom}\, f)$ and $z \in \partial f(y_0)$. Thus

$$f(y) \geq f(y_0) + \langle z,\, y - y_0 \rangle \quad \forall\, y \in \text{dom}\, f$$

and $\phi(y; x) \to +\infty$ as $\|y\| \to \infty$. Since $\phi(\cdot\,; x)$ is coercive, existence of a minimizer follows from standard compactness argument.

Examp 1. Let $f(x) = \rho\|x\|_1$ for $x \in \mathbb{R}^m$. Then $f^*(z) = \delta_C(z)$ where $C = \partial f(0) = \{z \in \mathbb{R}^m \mid \|z\|_\infty \leq \rho\}$.

Proof. For any z s.t. $\|z\|_\infty \leq \rho$. We have

$$\langle z,\, x\rangle \leq \|z\|_\infty \|x\|_1 \leq \rho\|x\|_1 = f(x) \quad \forall\, x \in \mathbb{R}^m \;\; \Rightarrow \;\; z \in \partial f(0)$$

Conversely for any $v \in \partial f(0)$, it holds that $f(x) \geq \langle v,\, x\rangle \; \forall\, x \in \mathbb{R}^m$. Take $x = \mathrm{sign}(v_j)e_j$, we get $\rho \geq |v_j|$ for any j. Hence $\|v\|_\infty \leq \rho$. □

$$\begin{aligned} P_f(x) &= x - P_{f^*}(x) = x - \Pi_C(x) \\ &= \mathrm{sign}(x) \circ \max\{|x| - \rho, 0\} \quad \text{(soft thresholding)} \end{aligned}$$

Examp 2. Let $C \subset \mathcal{E}$ be closed and convex. For $f = \delta_C$,

$$P_{\delta_C}(x) = \mathrm{argmin}_{y\in\mathcal{E}}\Big\{\delta_C(y) + \frac{1}{2}\|y - x\|\Big\} = \mathrm{argmin}_{y\in C}\frac{1}{2}\|y - x\|^2 = \Pi_C(x)$$

Suppose $C = \mathbb{S}^n_+$, the cone of $n \times n$ symmetric PSD matrices. Then

$$\Pi_C(x) = Q\mathrm{Diag}(d_+)Q^T \quad \text{using spectral decomp. } x = Q\mathrm{Diag}(d)Q^T$$

Examp 3. Let $f(x) = \|x\|_* = \|\sigma(x)\|_1$ be the nuclear norm of $x \in \mathbb{R}^{m\times n}$. Then $f^*(z) = \delta_C(z)$ where $C := \partial f(0) = \{z \in \mathbb{R}^{m\times n} \mid \|\sigma(z)\|_\infty \leq 1\}$.

Proof. Based on von Neumann's trace ineq. $|\langle x, z\rangle| \leq \langle \sigma(x), \sigma(z)\rangle$. Let $x = U\text{Diag}(\sigma(x))V^*$ be its SVD. Then

$$\begin{aligned} P_f(x) &= x - P_{f^*}(x) = x - \Pi_C(x) \\ &= U\big(\text{Diag}(\sigma(x)) - \text{Diag}(\min\{\sigma(x), 1\})\big)V^* \end{aligned}$$

Examp 4. Let $f(x) = \|\sigma(x)\|_\infty$ be spectral norm of $x \in \mathbb{R}^{m\times n}$. Then $f^* = \delta_C$ where $C = \partial f(0) = \{z \in \mathbb{R}^{m\times n} \mid \|\sigma(z)\|_1 \leq 1\}$, and

$$\begin{aligned} P_f(x) &= x - P_{f^*}(x) = x - \Pi_C(x) \\ &= x - U\text{Diag}(\Pi_B(\sigma(x)))V^*, \quad B = \{\sigma \in \mathbb{R}^m \mid \|\sigma\|_1 \leq 1\} \end{aligned}$$

Note: $\Pi_B(\sigma)$ can be computed analytically with $O(m)$ operations.

Theorem 8

1. *P_f and $Q_f := I - P_f$ are firmly nonexpansive, i.e.,*

$$\|P_f(x) - P_f(y)\|^2 \leq \langle P_f(x) - P_f(y), x - y \rangle \quad \forall\, x, y \in \mathcal{E}$$

2. *M_f is cont. differentiable and $\nabla M_f(x) = x - P_f(x)\ \forall\, x \in \mathcal{E}$.*

Proof. (1) From the def. of $x^* := P_f(x)$, $y^* := P_f(y)$, we get

$$0 \in \partial f(x^*) + x^* - x \;\Leftrightarrow \bar{x} := x - x^* \in \partial f(x^*), \;\ldots$$

Monotonicity of ∂f implies that $\langle x^* - y^*, \bar{x} - \bar{y} \rangle \geq 0$.

Hence $\|x^* - y^*\|^2 \leq \langle x^* - y^*, x - y \rangle \leq \|x^* - y^*\| \|x - y\|$
$\Rightarrow \|x^* - y^*\| \leq \|x - y\|$.

Similarly, $\|\bar{x} - \bar{y}\|^2 \leq \langle x - y, \bar{x} - \bar{y} \rangle$. Hence $\|\bar{x} - \bar{y}\| \leq \|x - y\|$.

Proof. (2) Using the identity $\frac{1}{2}(\|b\|^2 - \|a\|^2) = \langle a,\, b-a\rangle + \frac{1}{2}\|b-a\|^2$,

$$M_f(y) - M_f(x) = f(y^*) - f(x^*) + \frac{1}{2}(\|\bar{y}\|^2 - \|\bar{x}\|^2)$$
$$= \quad f(y^*) - f(x^*) + \langle \bar{x},\, y - x\rangle - \langle \bar{x},\, y^* - x^*\rangle + \frac{1}{2}\|\bar{y} - \bar{x}\|^2$$

Since $\bar{x} \in \partial f(x^*)$, $0 \le f(y^*) - f(x^*) - \langle \bar{x},\, y^* - x^*\rangle$. Similarly $\bar{y} \in \partial f(y^*)$ implies that

$$f(y^*) - f(x^*) \le \langle \bar{y},\, y^* - x^*\rangle \le \langle \bar{x},\, y^* - x^*\rangle + \|x - y\|^2.$$

Thus

$$0 \le \frac{1}{2}\|\bar{y} - \bar{x}\| \le M_f(y) - M_f(x) - \langle \bar{x},\, y - x\rangle \le \frac{3}{2}\|y - x\|^2.$$

□

Proposition 9 (Property of P_f)

Let f be cpc on $\mathcal{E}$. For any $x \in \mathcal{E}$, it holds that

1. *Any $V \in \partial P_f(x)$ is self-adjoint*
2. $\langle Vd,\, d\rangle \geq \|Vd\|^2 \;\; \forall\, d \in \mathcal{E}$.

Proof. (1) Consider $\phi : \mathcal{E} \to \mathbb{R}$ defined by $\phi(y) = \frac{1}{2}\|y\|^2 - M_f(y)$. It is continuously differentiable with $\nabla\phi(y) = P_f(y)$. Hence $(\nabla\phi)'(y)$ is self-adjoint if it exists. Thus any element of $\partial_B P_f(x)$ and that in $\partial P_f(x) = \mathrm{conv}(\partial_B P_f(x))$ is seld-adjoint.

For locally Lipschitz fun. $F : \mathcal{E} \to \mathcal{X}$ (finite-dim. Euclidean space),

$$\partial_B F(x) = \Big\{ \lim_{\mathcal{D}_F \ni z \to x} F'(z) \Big\}, \quad \partial F(x) = \mathrm{conv}(\partial_B F(x)).$$

(2) Let $d \in \mathcal{E}$ and $z \in \mathcal{D}_{P_f}$. Since P_f is nonexpansive,

$$\begin{aligned} & \langle P_f(z+td) - P_f(z),\, td\rangle \geq \|P_f(z+td) - P_f(z)\|^2 \\ \Rightarrow\quad & \langle tP_f'(z)d,\, td\rangle + o(t^2) \geq \|tP_f'(z)d + o(t)\|^2 \\ \Rightarrow\quad & \langle P_f'(z)d,\, d\rangle \geq \|P_f'(z)d\|^2 \quad \text{by letting } t \to 0 \end{aligned} \tag{1}$$

Let $V \in \partial P_f(x)$. There exists $V_i \in \partial_B P_f(x^i)$ such that

$$V = \sum_{i=1}^m \lambda_i V_i, \quad \sum_{i=1}^m \lambda_i = 1,\ \lambda_i \geq 0.$$

For each integer $k \geq 1$, $\exists\ x^{i_k}$ such that $\|x^{i_k} - x\| \leq k^{-1}$ and

$$\|P_f'(x^{i_k}) - V_i\| \leq k^{-1}$$

By (1), $\langle P_f'(x^{i_k})d,\, d\rangle \geq \|P_f'(x^{i_k})d\|^2$. Taking limits give

$$\langle V_i d,\, d\rangle \geq \|V_i d\|^2.$$

Thus

$$\begin{aligned} \langle Vd,\, d\rangle = \sum_{i=1}^m \lambda_i \langle V_i d,\, d\rangle &\geq \sum_{i=1}^m \lambda_i \|V_i d\|^2 \\ &\geq \|\sum_{i=1}^m \lambda_i V_i d\|^2 = \|Vd\|^2. \end{aligned}$$

Moreau decomposition

Theorem 10 (Moreau decomp.)

Let f be cpc and f^* be its conjugate. Then

$$
\begin{aligned}
x &= P_f(x) + P_{f^*}(x) \quad \forall\, x \in \mathcal{E} \\
\frac{1}{2}\|x\|^2 &= M_f(x) + M_{f^*}(x)
\end{aligned}
$$

For any $t > 0$,

$$
\begin{aligned}
x &= P_{tf}(x) + tP_{f^*/t}(x/t) \quad \forall\, x \in \mathcal{E} \\
\frac{1}{2}\|x\|^2 &= M_{tf}(x) + t^2 M_{f^*/t}(x/t)
\end{aligned}
$$

Proof. Let $z^* = x - x^*$ with $x^* = P_f(x)$. Now

$$
x^* = P_f(x) \;\Leftrightarrow\; x - x^* \in \partial f(x^*) \;\Leftrightarrow\; x^* \in \partial f^*(x - x^*)
$$

$$
\Leftrightarrow\; x - z^* \in \partial f(z^*) \;\Leftrightarrow\; z^* = P_{f^*}(x).
$$

Thus $x = x^* + z^* = P_f(x) + P_{f^*}(x)$.

$$\begin{aligned} M_f(x) + M_{f^*}(x) &= \min_{s,t} \left\{ \frac{1}{2}\|s-x\|^2 + \frac{1}{2}\|t-x\|^2 + f(s) + f^*(t) \right\} \\ &\geq \min_{s,t} \left\{ \frac{1}{2}\|s-x\|^2 + \frac{1}{2}\|t-x\|^2 + \langle s,\, t\rangle \right\} \\ &= \min_{s,t} \left\{ \frac{1}{2}\|s+t-x\|^2 + \frac{1}{2}\|x\|^2 \right\} \;\geq\; \frac{1}{2}\|x\|^2 \end{aligned}$$

Take $s = P_f(x)$. We have $x - s \in \partial f(s) \Leftrightarrow \langle s,\, x-s\rangle = f(s) + f^*(x-s)$. Let $t = x - s$. Then

$$\begin{aligned} M_f(x) + M_{f^*}(x) \;&\leq\; \frac{1}{2}\|s-x\|^2 + \frac{1}{2}\|t-x\|^2 + \langle s,\, t\rangle \\ = \;&\frac{1}{2}\|s+t-x\|^2 + \frac{1}{2}\|x\|^2 = \frac{1}{2}\|x\|^2. \end{aligned}$$

Thus we conclude that $M_f(x) + M_{f^*}(x) = \frac{1}{2}\|x\|^2$.

Directional derivatives

Let $X \in \mathcal{S}^n$ have the eigenvalue decomposition $X = \overline{P}\Lambda\overline{P}^T$. For a given index i, suppose

$$\lambda_{i-s_i}(X) > \lambda_{i-s_i+1} = \cdots = \lambda_i = \cdots = \lambda_{i+t_i} := \lambda^{(k)} > \lambda_{i+t_i+1}$$

$$J_k = \{1 \leq i \leq n \mid \lambda_i = \lambda^{(k)}\}$$

where $\lambda^{(k)}$'s denote the distinct eigenvalues of X.

Proposition 11 (Lancaster, Numerische Mathematik, 64)

Given $i \in J_k$, for any $\mathcal{S}^n \ni H \to 0$, we have

$$\lambda_i(X+H) - \lambda_i(X) - \lambda_{s_i}(\overline{P}_{J_k}^T H \overline{P}_{J_k}) = O(\|H\|^2).$$

Hence the eigenvalue function $\lambda_i(\cdot)$ is directionally differentiable at X with $\lambda_i'(X;H) = \lambda_{s_i}(\overline{P}_{J_k}^T H \overline{P}_{J_k})$.

Let $f : \mathbb{R} \to \mathbb{R}$ be a scalar function. Suppose $X \in \mathbb{S}^n$ has spectral decomp. $X = \overline{P}\Lambda\overline{P}^T$. The Löwner spectral operator is defined by

$$F(X) := \overline{P}\mathrm{Diag}(f(\lambda_1), \ldots, f(\lambda_n))\overline{P}^T.$$

Proposition 12 (Donoghue 74; Bhatia p.124)

The Löwner operator $F(\cdot)$ is (cont.) differentiable at X iff $f(\cdot)$ is (cont.) differentiable at each $\lambda_i(X)$. In this case, the Fréchet derivative is given by

$$F'(X)H = \overline{P}\left[f^{[1]}(\Lambda) \circ (\overline{P}^T H \overline{P})\right]\overline{P}^T \quad \forall\, H \in \mathcal{S}^n.$$

Here given $D = \mathrm{Diag}(d)$, $f^{[1]}(D) \in \mathbb{S}^n$ is defined by

$$(f^{[1]}(D))_{ij} = \begin{cases} \frac{f(d_i)-f(d_j)}{d_i-d_j} & \text{if } d_i \neq d_j \\ f'(d_i) & \text{if } d_i = d_j \end{cases}$$

Suppose $f(t) := t^+ = \max\{0, t\}$, and

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0 > \lambda_{r+1} \geq \cdots \geq \lambda_n$$

$$F(X) = \overline{P}\mathrm{Diag}(\lambda_1^+, \ldots, \lambda_n^+)\overline{P}^T \;=\; \overline{P}\mathrm{Diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)\overline{P}^T$$

$$F'(X)H = \overline{P}\Big[\Omega \circ (\overline{P}^T H \overline{P})\Big]\overline{P}^T$$

$$\Omega_{ij} = \begin{cases} 1 & \text{if } i, j = 1, \ldots, r \\ \frac{\lambda_i}{\lambda_i + |\lambda_j|} & \text{if } i = 1, \ldots, r,\ j = r+1, \ldots, n \\ 0 & \text{if } i, j = r+1, \ldots, n \end{cases}$$

$$\Omega = \begin{bmatrix} \mathbf{1}_r \mathbf{1}_r^T & \Omega_{12} \\ \Omega_{12}^T & \mathbf{0}_{(n-r)\times(n-r)} \end{bmatrix}$$

Efficient computation of $F'(X)H$ should exploit the $\mathbf{0}$ block and the all ones block.

Let $X \in \mathbb{R}^{m\times n}$ $(m \le n)$ has SVD

$$X = \overline{U}\,[\Sigma \;\; 0]\,\overline{V}^T = \overline{U}\,[\Sigma \;\; 0]\,\left[\overline{V}_1 \;\; \overline{V}_2\right]^T = \overline{U}\Sigma(X)\overline{V}_1^T$$

$\overline{U} \in \mathcal{O}^m$, $\overline{V} = \left[\overline{V}_1 \;\; \overline{V}_2\right] \in \mathcal{O}^n$ with $\overline{V}_1 \in \Re^{n\times m}$, $\overline{V}_2 \in \Re^{n\times(n-m)}$. Given an index i, suppose

$$\sigma_{i-s_i} > \sigma_{i-s_i+1} = \ldots = \sigma_i = \ldots = \sigma_{i+t_i} := \sigma^{(k)} > \sigma_{i+t_i+1}$$

where $\sigma^{(k)}$'s denote the distinct positive singular values of X. Define

$$a_k = \{1 \le i \le m \mid \sigma_i = \sigma^{(k)} > 0\}, \quad b = \{1 \le i \le m \mid \sigma_i = 0\}$$

Proposition 13 (Lewis and Sendov: Set-Valued Analysis, 2005)

For any $\Re^{m\times n} \ni H \to 0$*, we have for* $i = 1, \ldots, m$

$$\sigma_i(X + H) - \sigma_i(X) - \sigma_i'(X; H) = O(\|H\|^2)$$

$$\sigma_i'(X; H) = \begin{cases} \lambda_{s_i}\left(\mathrm{Sym}(\overline{U}_{a_k}^T H \overline{V}_{a_k})\right) & \textit{if } i \in a_k \\ \sigma_{s_i}\left(\overline{U}_b^T H \left[\overline{V}_b \; \overline{V}_2\right]\right) & \textit{if } i \in b \end{cases}$$

Proof. Let $g : \mathbb{R} \to \mathbb{R}$ be a scalar function s.t. $g(t) = -g(-t)$. The Löwner's singular value operator at X is defined by

$$G(X) := \overline{U}\,[g(\Sigma) \quad 0]\,\overline{V}^T = \sum_{i=1}^m g(\sigma_i)\,\bar{u}_i \bar{v}_i^T \qquad (2)$$

where $g(\Sigma) := \mathrm{Diag}(g(\sigma_1), \ldots, g(\sigma_m))$. It is known that

$$\mathcal{B}(G(X)) := \begin{bmatrix} 0 & G(X) \\ G(X)^T & 0 \end{bmatrix} = \overline{P} \begin{bmatrix} g(\Sigma) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -g(\Sigma) \end{bmatrix} \overline{P}^T = \mathcal{G}(\mathcal{B}(X))$$

where $\overline{P} \in \mathcal{O}^{m+n}$ is given by

$$\overline{P} = \frac{1}{\sqrt{2}} \begin{bmatrix} \overline{U} & 0 & \overline{U} \\ \overline{V}_1 & \sqrt{2}\,\overline{V}_2 & -\overline{V}_1 \end{bmatrix}.$$

Thus differential properties of $G(X)$ can be deduced from those of the spectral operator of $\mathcal{B}(X)$. Therefore, if $\mathcal{G}(\cdot)$ is (cont.) differentiable at $\mathcal{B}(X)$, $G(\cdot)$ is also (cont.) differentiable at X with

$$\mathcal{B}(G'(X)H) \;=\; \mathcal{G}'(\mathcal{B}(X))\mathcal{B}(H) \quad \forall\, H \in \Re^{m\times n} \qquad (3)$$

$$\begin{bmatrix} 0 & G'(X)H \\ (G'(X)H)^T & 0 \end{bmatrix} = \overline{P}\big(\Omega \circ (\overline{P}^T \mathcal{B}(H)\overline{P})\big)\overline{P}^T$$

$$\overline{P}^T \mathcal{B}(H)\overline{P} = \begin{bmatrix} H_1^s & \frac{1}{\sqrt{2}}H_2 & -H_1^a \\ & 0 & \frac{1}{\sqrt{2}}H_2^T \\ & & -H_1^s \end{bmatrix}, \quad \begin{array}{l} H_1 = U^T H V_1 \in \mathbb{R}^{m\times m} \\ H_2 = U^T H V_2 \in \mathbb{R}^{m\times(n-m)} \end{array}$$

where $H_1^s = \frac{1}{2}(H_1 + H_1^T)$, $H_1^a = \frac{1}{2}(H_1 - H_1^T)$.

$$\Omega = \begin{bmatrix} \Omega^- & \omega \mathbf{1}^T & \Omega^+ \\ & g'(0)\mathbf{1}\mathbf{1}^T & \mathbf{1}\omega^T \\ & & \Omega^- \end{bmatrix} \in \mathbb{S}^{m+n}, \quad \mathbf{1} \in \mathbb{R}^{n-m}$$

$$\Omega_{ij}^{\pm} = \begin{cases} \dfrac{g(\sigma_i) \pm g(\sigma_j)}{\sigma_i \pm \sigma_j} & \text{if } \sigma_i \pm \sigma_j \neq 0,\ i,j = 1,\ldots,m \\ g'(\sigma_i) & \text{otherwise} \end{cases}$$

$$\omega_i = \begin{cases} \dfrac{g(\sigma_i)}{\sigma_i} & \text{if } \sigma_i > 0,\ i = 1,\ldots,m \\ g'(0) & \text{if } \sigma_i = 0 \end{cases}$$

$$G'(X)H = U\Big[\Omega^- \circ H_1^a + \Omega^+ \circ H_1^a\Big]V_1^T + U\Big[(\omega \mathbf{1}^T) \circ H_2\Big]V_2^T$$

When $m \ll n$, expensive to compute $V_2 \in \mathbb{R}^{n \times (n-m)}$ explicitly. Fortunately, V_2 is not needed explicitly!

$$[(\omega \mathbf{1}^T) \circ H_2]V_2^T = \mathrm{Diag}(\omega) U^T H V_2 V_2^T$$

$$= \mathrm{Diag}(\omega) U^T H (I_n - V_1 V_1^T) = \mathrm{Diag}(\omega) U^T H - \mathrm{Diag}(\omega) H_1 V_1^T$$

Duality and examples

Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be real finite dim. real inner product spaces. Consider

$$\text{(CCP)} \qquad \min \left\{ h(\mathcal{A}x) + \theta(\mathcal{B}x) \right\}$$

$b \in \mathcal{Y}$ and $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$, $\mathcal{B} : \mathcal{X} \to \mathcal{Z}$ are linear maps
$h : \mathcal{Y} \to (\infty, \infty]$, $\theta : \mathcal{Z} \to (-\infty, \infty]$ are cpc fun.

Rewrite it as $\min \left\{ h(u) + \theta(v) \mid \mathcal{A}x - u = 0,\ \mathcal{B}x - v = 0 \right\}$.
Its Lagrangian fun. is

$$\mathcal{L}(x, u, v; y, \bar{y}) = h(u) + \theta(v) + \langle y,\, u - \mathcal{A}x \rangle + \langle \bar{y},\, v - \mathcal{B}x \rangle$$

$$\inf_{x,u,v} \mathcal{L} = \begin{cases} \inf_u \{ h(u) + \langle y,\, u \rangle \} + \inf_v \{ \theta(v) + \langle \bar{y},\, v \rangle \} \\ +\ \inf_x \{ \langle -x,\, \mathcal{A}^* y + \mathcal{B}^* \bar{y} \rangle \} \end{cases}$$

The dual, defined by $\max_{y \in \mathcal{Y},\, \bar{y} \in \mathcal{Z}} \left\{ \inf_{x,u,v} \mathcal{L}(x, u, v; y, \bar{y}) \right\}$, is

$$\max_{y \in \mathcal{Y},\, \bar{y} \in \mathcal{Z}} \left\{ -h^*(-y) - \theta^*(-\bar{y}) \mid \mathcal{A}^* y + \mathcal{B}^* \bar{y} = 0 \right\}$$

Duality: TV-norm regularized regression

$$\min\left\{\frac{1}{2}\|Ax-b\|^2+\lambda\|Bx\|_1\right\},\quad B=\begin{pmatrix}1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1\end{pmatrix}$$

Here $h(u)=\frac{1}{2}\|u-b\|^2$, $\theta(v)=\lambda\|v\|_1$. Now

$$h^*(-y)=\frac{1}{2}\|y\|^2-\langle y,\,b\rangle,\quad \theta^*(-\bar{y})=\delta_C(\bar{y}),\quad C=\{\bar{y}\mid\|\bar{y}\|_\infty\le\lambda\}.$$

Its dual problem is given by

$$\max\left\{-\frac{1}{2}\|y\|^2+\langle y,\,b\rangle-\delta_C(\bar{y})\mid A^*y+B^*\bar{y}=0\right\}$$

$$\equiv\quad \min\left\{\frac{1}{2}\|y\|^2-\langle y,\,b\rangle\mid A^*y+B^*\bar{y}=0,\ \|\bar{y}\|_\infty\le\lambda\right\}$$

Let Q be a closed convex cone in $\mathbb{R}^m$, and $\mathcal{X} = \mathbb{S}^n$.

$$\text{(SDP)} \quad \min \left\{ \langle c,\, x \rangle \mid \mathcal{A}x - b \in Q, x \in \mathbb{S}^n_+ \right\}$$

Let $\theta(x) = \langle c,\, x\rangle + \delta_{\mathbb{S}^n_+}(x)$, $h(u) = \delta_{b+Q}(u)$. The above SDP can be written as

$$\min \left\{ \theta(x) + h(\mathcal{A}x) \mid x \in \mathcal{X} \right\}$$

Its dual is given by

$$\max \left\{ -h^*(-y) - \theta^*(-\bar{y}) \mid \mathcal{A}^* y + \bar{y} = 0 \right\}$$

Now $h^*(-y) = -\langle y,\, b\rangle + \delta_{Q^*}(y)$ and $\theta^*(-\bar{y}) = \begin{cases} 0 & \text{if } c + \bar{y} \in \mathbb{S}^n_+ \\ \infty & \text{otherwise} \end{cases}$

More explicitly, the dual is

$$\max \left\{ \langle b,\, y\rangle \mid \mathcal{A}^* y + z - c = 0,\ z \in \mathbb{S}^n_+,\ y \in Q^* \right\}$$

Let $\mathcal{N}^n = \{x \in \mathbb{S}^n \mid x \geq\}$ (cone of nonnegative matrices).

(DNN) $$\min \Big\{ \langle c,\, x\rangle \mid \bar{\mathcal{A}}x - \bar{b} \in \bar{Q}, x \in \mathbb{S}^n_+, x \in \mathbb{N}^n \Big\}$$

$$= \min \Big\{ \langle c,\, x\rangle + \delta_{\mathbb{S}^n_+}(x) \mid \underbrace{\begin{bmatrix} \bar{\mathcal{A}} \\ I \end{bmatrix}}_{\mathcal{A}} x - \underbrace{\begin{bmatrix} b \\ 0 \end{bmatrix}}_{b} \in Q := \bar{Q} \times \mathbb{N}^n \Big\}$$

Let $\theta(x) = \langle c,\, x\rangle + \delta_{\mathbb{S}^n_+}(x)$, $h(u) = \delta_{b+Q}(u)$. The above DNN SDP can be written as in the form of (CCP) as

$$\min \Big\{ \theta(x) + h(\mathcal{A}x) \mid x \in \mathcal{X} \Big\}$$

Its dual is given by $\max \big\{ -h^*(-y) - \theta^*(-\bar{y}) \mid \mathcal{A}^* y + \bar{y} = 0 \big\}$. Now for $y = (y_1; y_2)$, $h^*(-y) = -\langle y,\, b\rangle + \delta_{Q^*}(y) = -\langle y_1,\, \bar{b}\rangle + \delta_{\bar{Q}^*}(y_1) + \delta_{\mathbb{N}^n}(y_2)$. The dual is explicitly given as

$$\max \Big\{ \langle \bar{b},\, y_1\rangle \mid \bar{\mathcal{A}}^* y_1 + y_2 + z - c = 0,\; y_1 \in \bar{Q}^*,\; y_2 \in \mathbb{N}^n,\; z \in \mathbb{S}^n_+ \Big\}$$

Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{U}$ be real finite dim. inner product spaces. Consider

$$\begin{aligned} \min \quad & \theta(x) + \tfrac{1}{2}\langle x,\, \mathcal{Q}x\rangle + \langle c,\, x\rangle + \phi(u) + \tfrac{1}{2}\langle u,\, \mathcal{P}u\rangle + \langle d,\, u\rangle \\ \text{s.t.} \quad & \mathcal{A}x + \mathcal{B}u = b \end{aligned}$$

$\mathcal{Q}$ and $\mathcal{P}$ are self-adjoint PSD linear operators on $\mathcal{X}$ and $\mathcal{U}$, resp.
$\mathcal{A} : \mathcal{X} \to \mathcal{Y}$, $\mathcal{B} : \mathcal{U} \to \mathcal{Y}$ are linear maps
$c \in \mathcal{X}$, $d \in \mathcal{U}$, $b \in \mathcal{Y}$ are given data
$\theta : \mathcal{X} \to (-\infty, \infty]$, $\phi : \mathcal{U} \to (-\infty, \infty]$ are cpc functions

$$\begin{aligned} \min \quad & \theta(\bar{x}) + \tfrac{1}{2}\langle x,\, \mathcal{Q}x\rangle + \langle c,\, x\rangle + \phi(\bar{u}) + \tfrac{1}{2}\langle u,\, \mathcal{P}u\rangle + \langle d,\, u\rangle \\ \text{s.t.} \quad & \mathcal{A}x + \mathcal{B}u = b, \quad \bar{x} - x = 0, \quad \bar{u} - u = 0 \end{aligned} \tag{4}$$

The Lagrangian function is

$$\mathcal{L}(x, u, \bar{x}, \bar{u}; y, s, z) = \left\{ \begin{array}{l} \theta(\bar{x}) + \frac{1}{2}\langle x,\ \mathcal{Q}x\rangle + \langle c,\ x\rangle + \langle y,\ b - \mathcal{A}x - \mathcal{B}u\rangle \\ +\phi(\bar{u}) + \frac{1}{2}\langle u,\ \mathcal{P}u\rangle + \langle d,\ u\rangle + \langle s,\ \bar{x} - x\rangle + \langle z,\ \bar{u} - u\rangle \end{array} \right.$$

The dual of (4) is defined by

$$\max_{y\in\mathcal{Y},\, s\in\mathcal{X},\, z\in\mathcal{U}} \Big\{ \inf_{x,\, u, \bar{x}, \bar{u}} \mathcal{L}(u, x, \bar{x}, \bar{u}; y, s, z) \Big\}$$

$$= \max \left\{ \begin{array}{l} \langle y,\ b\rangle + \inf_{\bar{x}} \{\theta(\bar{x}) + \langle s,\ \bar{x}\rangle\} + \inf_{\bar{u}} \{\phi(\bar{u}) + \langle z,\ \bar{u}\rangle\} \\ + \inf_{x\in\mathcal{X}} \left(\frac{1}{2}\langle x,\ \mathcal{Q}x\rangle - \langle \mathcal{A}^* y + s - c,\ x\rangle\right) \\ + \inf_{u\in\mathcal{U}} \left(\frac{1}{2}\langle u,\ \mathcal{P}u\rangle - \langle \mathcal{B}^* y + z - d,\ u\rangle\right) \end{array} \right\}$$

$$= \left\{ \begin{array}{ll} \max & \langle b,\ y\rangle - \theta^*(-s) - \phi^*(-z) - \frac{1}{2}\langle w,\ \mathcal{Q}w\rangle - \frac{1}{2}\langle v,\ \mathcal{P}v\rangle \\ \text{s.t} & -\mathcal{Q}w + \mathcal{A}^* y + s - c = 0,\ -\mathcal{P}v + \mathcal{B}^* y + z - d = 0 \\ & \quad w \in \mathcal{W},\ v \in \mathcal{V}. \end{array} \right. \quad (5)$$

where $\mathcal{W}$ is any subspace containing $\mathrm{Range}(\mathcal{Q})$, and $\mathcal{V}$ is any subspace containing $\mathrm{Range}(\mathcal{P})$.

$$\inf_{x\in\mathcal{X}} \left\{\tfrac{1}{2}\langle x,\, \mathcal{Q}x\rangle - \langle h,\, x\rangle\right\} = \begin{cases} -\infty & \text{if } h \notin \mathrm{Range}(\mathcal{Q}) \\ -\frac{1}{2}\langle w,\, Qw\rangle & \text{if } h = \mathcal{Q}w \end{cases}$$

The latter is achieved for any x s.t.

$$\mathcal{Q}(x - w) = 0 \;\Leftrightarrow\; x \in w + \mathrm{Null}(\mathcal{Q}).$$

In (5), in order to ensure that the primal optimal x can be recovered, w must at least be in a subspace containing $\mathrm{Range}(\mathcal{Q})$.

Convex composite quadratic conic programming:

$$\text{(QCP)} \quad \min \quad \theta(x) + \frac{1}{2}\langle x,\, \mathcal{Q}x\rangle + \langle c,\, x\rangle$$
$$\text{s.t.} \quad \mathcal{A}x - b \in \mathcal{C}, \quad x \in \mathcal{K},$$

where $\theta : \mathcal{X} \to (-\infty, +\infty]$ is a cpc fun.
$\mathcal{C} \subseteq \mathcal{Y}$, $\mathcal{K} \subseteq \mathcal{X}$ are closed convex cones
(QCP) obviously includes convex QP in $\mathbb{R}^n$ as a special case.

By introducing the var. $\bar{u} = \mathcal{A}x - b$ and $u - x = 0$, can rewrite (QCP) as

$$\begin{aligned} \min \quad & \delta_{\mathcal{K}}(x) + \frac{1}{2}\langle x,\, \mathcal{Q}x\rangle + \langle c,\, x\rangle + \theta(u) + \delta_{\mathcal{C}}(\bar{u}) \\ \text{s.t.} \quad & \begin{pmatrix} \mathcal{A} \\ I \end{pmatrix} x + \underbrace{\begin{pmatrix} -I & 0 \\ 0 & -I \end{pmatrix}}_{\mathcal{B}} \begin{pmatrix} \bar{u} \\ u \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix} \end{aligned} \tag{6}$$

Using (5), the dual of (6) is

$$\begin{array}{rl} \max & \langle b,\ y\rangle - \frac{1}{2}\langle w,\ \mathcal{Q}w\rangle - \delta^*_{\mathcal{K}}(-s) - \theta^*(-z) - \delta^*_{\mathcal{C}}(-\bar{z}) \\ \text{s.t.} & -\mathcal{Q}w + (\mathcal{A}^*,\ I)\begin{pmatrix} y \\ \bar{y} \end{pmatrix} + s = c \\ & \begin{pmatrix} -I & 0 \\ 0 & -I \end{pmatrix}\begin{pmatrix} y \\ \bar{y} \end{pmatrix} + \begin{pmatrix} \bar{z} \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{array}$$

So $\bar{z} = y$, $\bar{y} = z$. Hence

$$\begin{array}{rl} \max & \langle b,\ y\rangle - \frac{1}{2}\langle w,\ \mathcal{Q}w\rangle - \delta^*_{\mathcal{K}}(-s) - \theta^*(-z) - \delta^*_{\mathcal{C}}(-y) \\ \text{s.t.} & -\mathcal{Q}w + \mathcal{A}^* y + z + s = c, \quad \mathcal{Q} \in \mathcal{W} \end{array}$$

Noting that $\delta^*_{\mathcal{K}}(-s) = \delta_{\mathcal{K}^*}(s)$ and $\delta^*_{\mathcal{C}}(-y) = \delta_{\mathcal{C}^*}(y)$, we get

$$\begin{array}{rl} \max & \langle b,\ y\rangle - \frac{1}{2}\langle w,\ \mathcal{Q}w\rangle - \theta^*(-z) \\ \text{s.t.} & -\mathcal{Q}w + \mathcal{A}^* y + z + s = c, \quad \mathcal{Q} \in \mathcal{W} \\ & s \in \mathcal{K}^*, \quad y \in \mathcal{C}^* \end{array}$$

Examp. (QSDP with bound constraints)

$$\begin{array}{ll} \min & \frac{1}{2}\langle X,\, \mathcal{Q}X\rangle + \langle C,\, X\rangle \\ \text{s.t.} & \mathcal{A}_E X = b_E, \quad \mathcal{A}_I X \geq b_I, \quad X \in \mathbb{S}^n_+,\ X \in \mathcal{P} \end{array}$$

where $\mathcal{P} = \{X \in \mathbb{S}^n \mid L \leq X \leq U\}$ is a simple closed convex set.

$$\begin{array}{ll} \min & \delta_{\mathcal{P}}(X) + \frac{1}{2}\langle X,\, \mathcal{Q}X\rangle + \langle C,\, X\rangle \\ \text{s.t.} & \underbrace{\begin{bmatrix} \mathcal{A}_E \\ \mathcal{A}_I \end{bmatrix}}_{\mathcal{A}} X - \underbrace{\begin{bmatrix} b_E \\ b_I \end{bmatrix}}_{b} \in \mathcal{C} := \{\mathbf{0}_{m_E}\} \times \mathbb{R}^{m_I}_+, \quad X \in \mathcal{K} = \mathbb{S}^n_+ \end{array}$$

$$\begin{array}{ll} \max & -\delta^*_{\mathcal{P}}(-Z) - \frac{1}{2}\langle W,\, \mathcal{Q}W\rangle + \langle b_E,\, y_E\rangle + \langle b_I,\, y_I\rangle \\ \text{s.t.} & -\mathcal{Q}W + \mathcal{A}^*_E y_E + \mathcal{A}^*_I y_I + S + Z = C \\ & S \in \mathcal{K}^* = \mathbb{S}^n_+, \quad W \in \mathcal{W}, \quad y = \begin{bmatrix} y_E \\ y_I \end{bmatrix} \in \mathcal{C}^* = \mathbb{R}^{m_E} \times \mathbb{R}^{m_I}_+ \end{array}$$

Examp. (Regularized matrix least squares problem)

$$\begin{array}{ll} \min\limits_{X\in\mathbb{R}^{m\times n}} & \frac{1}{2}\|\mathcal{B}X - d\|^2 + \rho\|X\|_* \\ \text{s.t.} & \mathcal{A}X = b, \quad X \in \mathcal{P} = \{X \mid \|X_\Omega\|_\infty \le \alpha\} \end{array}$$

Let $\mathcal{Q} = \mathcal{B}^*\mathcal{B}$, $C = -\mathcal{B}^*d$, $\theta(X) = \rho\|X\|_*$, $\phi(X) = \delta_{\mathcal{P}}(X)$. We get

$$\begin{array}{ll} \min & \frac{1}{2}\langle X,\, \mathcal{Q}X\rangle + \langle C,\, X\rangle + \frac{1}{2}\|d\|^2 + \theta(X) + \phi(X) \\ \text{s.t.} & \mathcal{A}X = b \end{array}$$

Its dual is given by

$$\begin{array}{ll} \max & -\frac{1}{2}\langle W,\, \mathcal{Q}W\rangle + \langle b,\, y\rangle - \theta^*(-S) - \phi^*(-Z) + \frac{1}{2}\|d\|^2 \\ \text{s.t.} & -\mathcal{Q}W + \mathcal{A}^*y + S + Z = C, \quad W \in \mathbb{R}^{m\times n} \end{array}$$

$\theta^*(S) = \delta_{B_\rho}(S)$, $B_\rho = \{\|S\|_2 \le \rho\}$. Let $\xi = d - \mathcal{B}W$.

$$\begin{array}{ll} \max & -\frac{1}{2}\|\xi\|^2 + \langle d,\, \xi\rangle + \langle b,\, y\rangle - \phi^*(-Z) \\ \text{s.t.} & \mathcal{B}^*\xi + \mathcal{A}^*y + S + Z = 0, \quad \|S\|_2 \le \rho \end{array}$$

Consider the separable convex programming

$$\begin{array}{ll} \min & \theta_1(y_1) + \cdots + \theta_p(y_p) - \langle b_1,\, y_1\rangle - \cdots - \langle b_p,\, y_p\rangle \\ \text{s.t.} & \mathcal{A}_1^* y_1 + \cdots + \mathcal{A}_p^* y_p = c \end{array}$$

where $\theta_i : \mathcal{Y}_i \to (\infty, \infty]$ are cpc fun. Show that its dual is

$$\max\Big\{ -\langle c,\, x\rangle - \theta_1^*(b_1 - \mathcal{A}_1 x) - \cdots - \theta_p^*(b_p - \mathcal{A}_p x)\Big\}$$

$$\equiv \left\{ \begin{array}{ll} -\min & \Big\{\langle c,\, x\rangle + \theta_1^*(s_1) + \cdots + \theta_p^*(s_p) \\ \text{s.t.} & \begin{pmatrix} \mathcal{A}_1 \\ \vdots \\ \mathcal{A}_p \end{pmatrix} x + \begin{pmatrix} s_1 \\ \vdots \\ s_p \end{pmatrix} = b \end{array} \right.$$

A problem with 2 blocks: x and $(s_1; \cdots; s_p)$

Examples fit the following general convex composite model:

$$\min_{x\in\mathcal{X},y\in\mathcal{Y}} \{\theta(x_1) + f(x_1,\ldots,x_m) + \phi(y_1) + g(y_1,\ldots,y_n)\}$$

$$\text{s.t.} \quad \textstyle\sum_{i=1}^m \mathcal{A}_i^* x_i + \sum_{j=1}^n \mathcal{B}_j^* y_j = c \quad \rightarrow \; \mathcal{A}^* x + \mathcal{B}^* y = c$$

$x = (x_1,\ldots,x_m) \in \mathcal{X} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_m$

$y = (y_1,\ldots,y_n) \in \mathcal{Y} := \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_n$

$\theta : \mathcal{X}_1 \to (-\infty,\infty]$, $\phi : \mathcal{Y}_1 \to (-\infty,\infty]$ are proper closed convex

$f : \mathcal{X} \to \Re$, $g : \mathcal{Y} \to \Re$ are convex quadratic functions

$\mathcal{A}_i : \mathcal{Z} \to \mathcal{X}_i$, $\mathcal{B}_j : \mathcal{Z} \to \mathcal{Y}_i$ are given linear maps.

(CCQP) $\min_{x\in\mathcal{X},y\in\mathcal{Y}} \left\{\theta(x_1) + f(x) + \phi(y_1) + g(y) \mid \mathcal{A}^* x + \mathcal{B}^* y = c\right\}$

Least squares conic programming: SGS, SNCG, Danskin-type theorem

- C. Ding, *An Introduction to A Class of Matrix Optimization Problems*, PhD Thesis, National University of Singapore, 2012.
- X.D. Li, D.F. Sun, and K.C. Toh, *A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions*, arXiv:1409.2679, Math. Prog., to appear.
- H. Qi and D.F. Sun, *A quadratically convergent Newton method for computing the nearest correlation matrix*, SIAM J. Matrix Analysis and Applications, 28 (2006) 360–385.
- D.F. Sun and J. Sun, *Semismooth Matrix Valued Functions*, Mathematics of Operations Research, 27 (2002) 150–169.
- D.F. Sun, K.C. Toh, and L. Yang, *An efficient inexact ABCD method for least squares semidefinite programming*, SIAM J. on Optimization 26 (2016) 1072–1100.
- Xinyuan Zhao, Defeng Sun, and K. C. Toh, *A Newton-CG augmented Lagrangian method for semidefinite programming*, SIAM J. on Optimization, 20 (2010) 1737–1765.

Let $\mathbb{K}$ be a closed convex cone in a real finite-dim inner product space $\mathcal{X}$. Given $G \in \mathcal{X}$, consider

$$\min \left\{ \frac{1}{2}\|X - G\|^2 + p(X) \mid \mathcal{A}X = b \right\}$$

where $p(X) = \delta_{\mathbb{K}}(X)$, $\mathcal{A} : \mathcal{X} \to \mathbb{R}^m$ is a surjective linear map (hence $\mathcal{A}\mathcal{A}^*$ is nonsingular). Its dual is

$$-\min_{S,y} \left\{ \theta(S) + \frac{1}{2}\|S + \mathcal{A}^* y + G\|^2 - \langle b,\, y\rangle \right\} \;\; + \text{const}$$

$\theta(S) := p^*(-S) = \delta_{\mathbb{K}^*}(S)$. It has the form

$$\min_{S,y} \left\{ \theta(S) + q(S,y) \right\}$$

with quadratic function

$$q(S,y) = \langle (S;y),\, \mathcal{Q}(S;y)\rangle - \langle (S;y),\, (-G; b - \mathcal{A}G)\rangle,\; \mathcal{Q} \;\; = \begin{bmatrix} I & \mathcal{A}^* \\ \mathcal{A} & \mathcal{A}\mathcal{A}^* \end{bmatrix}$$

A symmetric (block) Gauss-Seidel decomposition theorem

Consider convex quadratic problem:

$$\min\left\{\theta(x_1) + q(x) \mid x = (x_1; x_2; \ldots; x_s) \in \mathcal{X} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_s\right\} \quad (7)$$

Convex quadratic function $q(x) := \frac{1}{2}\langle x, \mathcal{Q}x\rangle - \langle b, x\rangle$

Closed proper convex function $\theta : \mathcal{X}_1 \to (-\infty, +\infty]$

Consider the following block decomposition

$$\mathcal{Q}x \equiv \begin{pmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} & \cdots & \mathcal{Q}_{1s} \\ \mathcal{Q}_{12}^* & \mathcal{Q}_{22} & \cdots & \mathcal{Q}_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{Q}_{1s}^* & \mathcal{Q}_{2s}^* & \cdots & \mathcal{Q}_{ss} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_s \end{pmatrix}$$

Here, we further assume that $\mathcal{Q}_{ii} \succ 0 \; \forall\, i = 1, \ldots, s.$

Write $\mathcal{Q} = \mathcal{U}^* + \mathcal{D} + \mathcal{U}$
$\mathcal{D}$ = block diagonal part, $\mathcal{U}$ = strictly upper triangular part.

Define $\mathcal{Q}^{\mathsf{SGS}} : \mathcal{X} \to \mathcal{X}$

$$\mathcal{Q}^{\mathsf{SGS}} = \mathcal{U}\mathcal{D}^{-1}\mathcal{U}^* \quad \text{(symmetric Gauss-Seidel decomp.)}$$

Given $\bar{x} \in \mathcal{X}$, define

$$x^+ := \operatorname{argmin}_x \left\{ \theta(x_1) + q(x) + \frac{1}{2}\|x - \bar{x}\|^2_{\mathcal{Q}^{\mathsf{SGS}}} \right\} \tag{8}$$

Next theorem: can compute x^+ using one cycle of symmetric GS!

If $\theta(x_1)$ is absent, we get the classical block symmetric GS iteration.

Theorem 14 (Li-Sun-T.)

It holds that $\mathcal{Q} + \mathcal{Q}^{SGS} = (\mathcal{D}+\mathcal{U})\mathcal{D}^{-1}(\mathcal{D}+\mathcal{U}^*) \succ 0.$

Backward GS ($x_s \to x_{s-1} \to \cdots \to x_2$). *For* $i = s, \ldots, 2$, *compute*

$$\begin{aligned} x_i' &= \text{argmin}_{x_i}\ \theta(\bar{x}_1) + q(\bar{x}_{\leq i-1}, x_i, x'_{\geq i+1}) \\ &= \mathcal{Q}_{ii}^{-1}\Big(b_i - \textstyle\sum_{j=1}^{i-1} \mathcal{Q}_{ji}^* \bar{x}_j - \sum_{j=i+1}^{s} \mathcal{Q}_{ij} x_j'\Big) \quad \textit{(linear system)} \end{aligned}$$

Forward GS ($x_1 \to x_2 \to \cdots \to x_s$). *For* $i = 2, \ldots, s$, *compute*

$$\begin{aligned} x_1^+ &= \text{argmin}_{x_1}\ \theta(x_1) + q(x_1, x'_{\geq 2}) \quad \textit{(opt. prob. involving only } x_1\textit{)} \\ x_i^+ &= \text{argmin}_{x_i}\ \theta(x_1^+) + q(x^+_{\leq i-1}, x_i, x'_{\geq i+1}) \\ &= \mathcal{Q}_{ii}^{-1}(b_i - \textstyle\sum_{j=1}^{i-1} \mathcal{Q}_{ji}^* x_j^+ - \sum_{j=i+1}^{s} \mathcal{Q}_{ij} x_j') \quad \textit{(linear system)} \end{aligned}$$

Inexact computation is also allowed! So can use PCG to solve large linear systems.

Theorem 15 (Li-Sun-T.)

Backward GS: For $i = s, \ldots, 2$, compute

$$x_i' \;=\; \mathcal{Q}_{ii}^{-1}\big(b_i + \delta_i' - \textstyle\sum_{j=1}^{i-1} \mathcal{Q}_{ji}^* \bar{x}_j - \sum_{j=i+1}^{s} \mathcal{Q}_{ij} x_j'\big).$$

Forward GS: For $i = 2, \ldots, s$

$$x_1^+ \;=\; \text{argmin}_{x_1}\; \theta(x_1) + q(x_1, x_{\geq 2}') - \langle \delta_1^+, \, x_1 \rangle$$

$$x_i^+ \;=\; \mathcal{Q}_{ii}^{-1}(b_i + \delta_i^+ - \textstyle\sum_{j=1}^{i-1} \mathcal{Q}_{ji}^* x_j^+ - \sum_{j=i+1}^{s} \mathcal{Q}_{ij} x_j')$$

δ^+, δ' are error vectors. In this case, x^+ is the exact solution to a slightly perturbed proximal problem:

$$x^+ := \text{argmin}_x \left\{ \theta(x_1) + q(x) + \frac{1}{2}\|x - \bar{x}\|^2_{\mathcal{Q}\text{SGS}} - \langle x, \, \Delta(\delta', \delta^+) \rangle \right\}$$

$$\Delta(\delta', \delta^+) = \delta^+ + \mathcal{U}\mathcal{D}^{-1}(\delta^+ - \delta').$$

Cost saving

To avoid solving the ith problem in forward GS sweep for $i = 2, \ldots, s$, we may try to estimate x_i^+ using x_i', and the corresponding error vector δ_i^+ is given by

$$\delta_i^+ = \delta_i' + \sum_{j=1}^{i-1} \mathcal{Q}_{ji}^*(x_j^+ - \bar{x}_j).$$

One may accept such an approximate solution $x_i^+ = x_i'$ if the error vector satisfies the condition $\|\delta_i^+\| \leq c\|\delta_i'\|$ for some constant $c > 1$, say $c = 10$.

Proof. The optimality condition for (8) is

$$(\mathcal{Q}+\mathcal{Q}^{\mathsf{SGS}})x = b + \mathcal{Q}^{\mathsf{SGS}}\bar{x} - \gamma, \quad \gamma = \begin{pmatrix} \gamma_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \gamma_1 \in \partial\theta(x_1)$$

Since $\mathcal{D} \succ 0$, both $\mathcal{D}+\mathcal{U}$, $\mathcal{D}+\mathcal{U}^*$ are nonsingular. From the computation, we have

$$\begin{aligned} (\mathcal{D}+\mathcal{U})x' &= b - \mathcal{U}^*\bar{x} \\ (\mathcal{D}+\mathcal{U}^*)x &= b - \mathcal{U}x' - \gamma = \mathcal{D}x' + \mathcal{U}^*\bar{x} - \gamma \end{aligned}$$

Now one can verify that

$$(\mathcal{Q}+\mathcal{Q}^{\mathsf{SGS}})x = b + \mathcal{H}^{\mathsf{SGS}}\bar{x} - (\mathcal{D}+\mathcal{U})\mathcal{D}^{-1}\gamma$$

By the special form of γ, we have $(\mathcal{D}+\mathcal{U})\mathcal{D}^{-1}\gamma = \gamma$. The optimality condition is satisfied. □

One can use an inexact APG method to solve (7). Consider

$$\min_{x\in\mathcal{X}} \{F(x) := p(x) + f(x)\}$$

with $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall\, x, y \in \mathcal{X}$.
$p : \mathcal{X} \to (-\infty, \infty]$ is a cpc fun.

Algorithm. Input $\bar{x}^1 = x^0 \in \mathrm{dom}\,(p)$, $t_1 = 1$. Iterate

1. Find an approximate minimizer

$$x^k \approx \mathrm{argmin}_{x\in\mathcal{X}}\Big\{p(x)+f(\bar{x}^k)+\langle\nabla f(\bar{x}^k),\, x-\bar{x}^k\rangle+\frac{1}{2}\langle x-\bar{x}^k,\, \mathcal{H}_k(x-\bar{x}^k)\rangle\Big\}$$

   where $\mathcal{H}_k \succ 0$ is an a priori given PSD linear operator s.t. the blue part majorizes $f(x)$ for all $x \in \mathcal{X}$

2. Compute $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$, $\bar{x}^{k+1} = x^k + \left(\frac{t_k - 1}{t_{k+1}}\right)(x^k - x^{k-1})$.

Note $t_k \approx k/2$ for k large.

Consider the following admissible condition

$$\nabla f(\bar{x}^k) + \mathcal{H}_k(x^k - \bar{x}^k) + \gamma^k =: \delta^k \quad \text{with } \|\mathcal{H}_k^{-1/2}\delta^k\| \le \frac{\epsilon_k}{\sqrt{2}t_k}$$

where $\gamma^k \in \partial p(x^k)$ = the subdifferential of p at x^k
$\{\epsilon_k\}$ is a nonnegative summable sequence.

Theorem 16 (Jiang-Sun-T. SIOPT 2012)

Let x^* be an optimal solution. Suppose the admissible condition holds and $\mathcal{H}_{k-1} \succeq \mathcal{H}_k \succ 0 \ \forall\, k$. Then

$$0 \le F(x^k) - F(x^*) \le \frac{4}{(k+1)^2}\left(\sqrt{\tau} + \bar{\epsilon}_k\right)^2$$

where $\tau = \frac{1}{2}\|x^0 - x^*\|^2_{\mathcal{H}_1}$, $\bar{\epsilon}_k = \sum_{j=1}^k \epsilon_j$.

For the problem (7): $\min\{\theta(x_1) + q(x) \mid x \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_s\}$.

We have $p(x) = \theta(x_1)$, $f(x) = q(x) = \frac{1}{2}\langle x, \mathcal{Q}x\rangle - \langle b, x\rangle$. Pick $\mathcal{H}_k = \mathcal{H} := \mathcal{Q} + \mathcal{Q}^{\mathsf{SGS}}$, the APG subproblem is given by

$$x^k \;\approx\; \mathrm{argmin}\Big\{\theta(x_1) + q(x) + \frac{1}{2}\|x - \bar{x}^k\|^2_{\mathcal{Q}^{\mathsf{SGS}}}\Big\}$$

which can be solved by one cycle of inexact SGS and

$$\nabla f(\bar{x}^k) + \mathcal{H}(x^k - \bar{x}^k) + \gamma^k = \Delta(\delta', \delta^+)$$

where $\gamma^k = (\gamma_1^k; 0; \cdots; 0)$ and $\gamma_1^k \in \partial\theta(x_1^k)$. The admissible cond. is satisfied if

$$\|\mathcal{H}^{-1/2}\Delta(\delta', \delta^+)\| \leq \frac{\epsilon_k}{\sqrt{2}t_k}$$

Given $G \in \mathcal{X}$, consider

$$\text{(P)} \quad \min \left\{ \frac{1}{2}\|X - G\|^2 + p(X) \mid \mathcal{A}X = b \right\}$$

where $p(X) = \delta_{\mathbb{K}}(X)$, $\mathcal{A} : \mathcal{X} \to \mathbb{R}^m$ is a surjective linear map (hence $\mathcal{A}\mathcal{A}^*$ is nonsingular). Its dual is

$$\text{(D)} \quad -\min_{S,y} \left\{ \theta(S) + \frac{1}{2}\|S + \mathcal{A}^* y + G\|^2 - \langle b,\, y\rangle \right\} \quad + \text{const}$$

$\theta(S) := p^*(-S) = \delta_{\mathbb{K}^*}(S)$. It has the form (7) with

$$q(S, y) \;=\; \langle (S; y),\, \mathcal{Q}(S; y)\rangle - \langle (S; y),\, (-G; b - \mathcal{A}G)\rangle$$

$$\mathcal{Q} \;=\; \begin{bmatrix} I & \mathcal{A}^* \\ \mathcal{A} & \mathcal{A}\mathcal{A}^* \end{bmatrix}, \quad \mathcal{Q}^{\mathsf{SGS}} = \begin{bmatrix} \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A} & 0 \\ 0 & 0 \end{bmatrix}$$

The APG subproblem is given by

$$(S^k, y^k) \;\approx\; \text{argmin}\Big\{\theta(S) + q(S,y) + \frac{1}{2}\|(S;y) - (\bar{S};\bar{y})^k\|^2_{\mathcal{Q}\text{SGS}}\Big\}$$

Inexact ABCD. Input $\bar{S}^1 = S^0 \succeq 0$, $t_1 = 1$. Iterate

1. Compute based on SGS:

$$\begin{aligned} y^{k+\frac{1}{2}} &\approx \text{argmin}_y\{\frac{1}{2}\|\bar{S}^k + \mathcal{A}^* y + G\|^2\} \\ S^k &= \text{argmin}_S\{\theta(S) + \frac{1}{2}\|S + \mathcal{A}^* y^{k+\frac{1}{2}} + G\|^2\} = \Pi_{\mathbb{K}^*}(-\mathcal{A}^* y^{k+\frac{1}{2}} - G) \\ y^{k+1} &\approx \text{argmin}_y\{\frac{1}{2}\|S^k + \mathcal{A}^* y + G\|^2\} \end{aligned}$$

2. Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$, $(\bar{S},\bar{y})^{k+1} = (S,y)^k + \Big(\frac{t_k - 1}{t_{k+1}}\Big)\big((S,y)^k - (S,y)^{k-1}\big)$.

One can attempt to directly solve the dual problem:

$$
\begin{aligned}
& \min_{S,y} \Big\{ \theta(S) + \frac{1}{2}\|S + \mathcal{A}^* y + G\|^2 - \langle b,\, y\rangle \Big\} \\
= \quad & \min_{y} \Big\{ -\langle b,\, y\rangle + \min_{S} \{ p^*(-S) + \frac{1}{2}\|S + \mathcal{A}^* y + G\|^2 \} \Big\} \\
= \quad & \min_{y} \Big\{ \Phi(y) := -\langle b,\, y\rangle + M_{p^*}(\mathcal{A}^* y + G) \Big\} \qquad (9)
\end{aligned}
$$

where $M_{p^*}(\cdot)$ is the MY regularization of p^*. Let $P_p(\cdot)$ be the proximal map for p. Its optimality condition is

$$
0 = \nabla\Phi(y) = -b + \mathcal{A}\nabla M_{p^*}(\mathcal{A}^* y + G) = -b + \mathcal{A}P_p(\mathcal{A}^* y + G)
$$

[Qi, Sun, SIMAX 2006] studied SNCG method for solving nearest correlation matrix problems where $\theta(S) = \Pi_{\mathbb{S}^n_+}$, $\mathcal{A}(X) = \mathrm{diag}(X)$, $b = \mathbf{1}$, and demonstrated its impressive efficiency!

Solve $\nabla\Phi(y) = -b + \mathcal{A}\Pi_{\mathbb{S}^n_+}(U) = 0, \quad U = \mathcal{A}^* y + G.$

$\nabla\Phi(y)$ is not differentiable, but is strongly semismooth [Sun-Sun, MOR 2002]. At current iterate y^l, solve a generalized Newton eq.

$$\mathcal{H}\Delta y \approx -\nabla\Phi(y^l), \quad \text{where } \mathcal{H}\Delta y = \mathcal{A}\Pi'_{\mathbb{S}^n_+}(U^l)[\mathcal{A}^*\Delta y] \qquad (10)$$

From eigenvalue decomp: $U^l = QDQ^T$ with $d_1 \geq \cdots \geq d_r \geq 0 > d_{r+1} \geq \cdots \geq d_n$, we choose

$$\Pi'_{\mathbb{S}^n_+}(U^l)[M] = Q(\Omega \circ (Q^T M Q))Q^T, \qquad (11)$$

where $\Omega_{ij} = (d_i^+ - d_j^+)/(d_i - d_j)$. For $\gamma = \{1, \ldots, r\}$ and $\bar{\gamma} = \{r+1, \ldots, n\}$, we have

$$\Omega = \begin{bmatrix} E_{\gamma\gamma} & \Omega_{\gamma\bar{\gamma}} \\ \Omega_{\bar{\gamma}\gamma} & 0 \end{bmatrix}$$ (we call such structure 2nd order sparsity)

The structure in Ω allows for efficient comp. of rhs of (11), and hence matrix-vector multiply for CG in solving (10)

Algorithm: Pick $\eta \in (0,1)$, $\tau \in (0,1]$, $\tau_1, \tau_2 \in (0,1)$ and $y^0 \in \mathbb{R}^m$, iterate the following steps.

1. Pick $\mathcal{V}_j \in \widehat{\partial}^2 \Phi(y^j)$, set $\epsilon_j = \tau_1 \min\{\tau_2, \|\nabla \Phi(y^j)\|\}$. Apply the PCG to compute an approx. solution d^j for

$$(\mathcal{V}_j + \epsilon_j I)d = -\nabla \Phi(y^j) \tag{12}$$

s.t. $\|(\mathcal{V}_j + \epsilon_j I)d^j + \nabla \Phi(y^j)\| \leq \eta_j := \min\{\eta, \|\nabla \Phi(y^j)^{1+\tau}\}$

2. Fix $\varsigma \in (0, 0.5)$, $\rho \in (0,1)$. Find smallest nonnegative integer s s.t.

$$\Phi(y^j + \rho^s d^j) \leq \Phi(y^j) + \varsigma \rho^s \langle \nabla \Phi(y^j), d^j \rangle.$$

Set $y^{j+1} := y^j + \rho^s d^j$.

$$\widehat{\partial}^2 \Phi(y) = \mathcal{A} \partial P_p(\mathcal{A}^* y + G)\mathcal{A}^*$$

By [Clarke p.75], $\partial^2 \Phi(y) H \subset \widehat{\partial}^2 \Phi(y) H$ for all $H \in \mathcal{X}$.

Assume that $P_p(\cdot)$ is strongly semismooth and $\{y \mid \Phi(y) \leq \Phi(y^0)\}$ is bounded.

Theorem 17

Let $\widehat{y}$ be an accum. point of the seq. $\{y^j\}$ generated by SNCG Algo. for solving (9) *s.t.*

$$\|\nabla\Phi(y^j) + (\mathcal{V}_j + \varepsilon_j I)\, d^j\| \leq \eta_j \quad \forall\, j. \tag{13}$$

Then $\widehat{y}$ is an optimal sol. to (9)*. If* **constraint nondegeneracy** *condition holds at $\widehat{X} = \mathcal{A}^*\widehat{y} + G$, then $\{y^j\}$ converges to $\widehat{y}$, and*

$$\|y^{j+1} - \widehat{y}\| = O(\|y^j - \widehat{y}\|^{1+\tau}).$$

Proof. By the Lip. continuity of P_p, easy to show that $\nabla\Phi(\widehat{y}) = 0$. Under constraint nondegeneracy, for any $\mathcal{V}_{\widehat{y}} \in \widehat{\partial}^2\Phi(\widehat{y})$, $\exists\ W_{\widehat{y}} \in \partial P_p(\widehat{X})$ s.t.

$$\mathcal{V}_{\widehat{y}} = \mathcal{A}W_{\widehat{y}}\mathcal{A}^* \succ \mathbf{0}.$$

Then, for j sufficiently large, $\{\|(\mathcal{V}_j + \varepsilon_j I)^{-1}\|\}$ is uniformly bounded.

For any $\mathcal{V}_j$, $\exists\ W_j \in \partial P_p(\mathcal{A}^* y^j + G))$ s.t. $\mathcal{V}_j = \mathcal{A} W_j \mathcal{A}^*$.

$$
\begin{aligned}
\|y^j + d^j - \widehat{y}\| \ &\leq \ \|y^j - \widehat{y} - (\mathcal{V}_j + \varepsilon_j I)^{-1} \nabla\Phi(y^j)\| \\
&\qquad + \|(\mathcal{V}_j + \varepsilon_j I)^{-1}\| \, \|\nabla\Phi(y^j) + (\mathcal{V}_j + \varepsilon_j I)\, d^j\| \\
&\leq \ \|(\mathcal{V}_j + \varepsilon_j I)^{-1}\| \Big(\|\nabla\Phi(y^j) - \nabla\Phi(\widehat{y}) - \mathcal{V}_j(y^j - \widehat{y})\| + \varepsilon_j \|y^j - \widehat{y}\| + \eta_j \Big) \\
&\leq \ O\Big(\|\mathcal{A}\| \, \|\Pi_{\mathbb{S}^n_+}(\mathcal{A}^* y^j + G) - \Pi_{\mathbb{S}^n_+}(\mathcal{A}^* \widehat{y} + G) - W_j(\mathcal{A}^*(y^j - \widehat{y}))\| \Big) \\
&\quad + O\big(\tau_1 \|\nabla\Phi(y^j)\| \|y^j - \widehat{y}\| + \|\nabla\Phi(y^j)\|^{1+\tau} \big) \\
&\leq \ O(\|\mathcal{A}^*(y^j - \hat{y})\|^2) \ + \ \cdots \quad \text{(since } P_p \text{ is strongly semismooth)} \\
&= \ O(\|y^j - \widehat{y}\|^{1+\tau})
\end{aligned}
$$

which implies that for all j sufficiently large,

$$
y^j - \widehat{y} = -d^j + O(\|d^j\|^{1+\tau}) \quad \text{and} \quad \|d^j\| \to 0.
$$

$$
\begin{aligned}
\langle \nabla \Phi(y^j),\, d^j \rangle + \langle d^j,\, (\mathcal{V}_j + \varepsilon_j I)\, d^j \rangle \;&\le\; \eta_j \|d^j\| \\
\le\;\; \|\nabla \Phi(y^j)\|^{1+\tau}\, \|d^j\| \;&=\; \|\nabla \Phi(y^j) - \nabla \Phi(\widehat{y})\|^{1+\tau} \|d^j\| \\
\le\;\; O(\|d^j\|^{2+\tau}) \quad &\text{for } j \text{ sufficiently large}
\end{aligned}
$$

which, together with the fact that $\|(\mathcal{V}_j + \varepsilon_j I)^{-1}\|$ is uniformly bounded, implies $\exists$ a constant $\delta > 0$ s.t.

$$-\langle \nabla \Phi(y^j),\, d^j \rangle \ge \delta \|d^j\|^2 \quad \text{for } j \text{ sufficiently large.}$$

Since $\nabla \Phi(\cdot)$ is (strongly) semismooth at $\widehat{y}$, from [Facchinei95, Thm 3.3 & Remark 3.4], we know that for $\varsigma \in (0, 1/2)$ and j sufficiently large,

$$\Phi(y^j + d^j) \le \Phi(y^j) + \varsigma \langle \nabla \Phi(y^j),\, d^j \rangle$$

Hence for j sufficiently large,

$$y^{j+1} = y^j + d^j.$$

Let $\mathcal{P} = \{X \in \mathbb{S}^n \mid L \leq X \leq U\}$. Consider

$$\begin{aligned} \min \quad & \tfrac{1}{2}\|X - G\|^2 \\ \text{s.t.} \quad & \mathcal{A}(X) = b, \; X \in \mathbb{S}^n_+, \; X \in \mathcal{P}. \end{aligned}$$

The dual is given by

$$\boxed{-\min\left\{\delta^*_{\mathcal{P}}(-Z) + \delta^*_{\mathcal{S}^n_+}(-S) + \frac{1}{2}\|Z + S + \mathcal{A}^* y + G\|^2 - \langle b, \, y\rangle\right\} + \text{const}}$$

It has the form:

$$\min\left\{\boldsymbol{F}(\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{y}) := \delta^*_{\mathcal{P}}(-Z) + \delta^*_{\mathcal{S}^n_+}(-S) + \phi(Z, S, y)\right\}$$

Here $\phi(Z, S, y)$ is convex quadratic fun.

Consider block vectors $x = (x_1, x_2, \ldots, x_s) \in \mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \cdots \times \mathcal{X}_s$

$$\begin{aligned} & \min\{p(x_1) + \varphi(z) + \phi(z, x) \mid z \in \mathcal{Z},\ x \in \mathcal{X}\} \\ = \quad & \min\{p(x_1) + f(x) \mid x \in \mathcal{X}\} \end{aligned}$$

where $p(\cdot)$, $\varphi(\cdot)$ are cpc functions, and

$$f(x) = \min_{z \in \mathcal{Z}} \{\varphi(z) + \phi(z, x)\}$$

$$z(x) = \operatorname{argmin}\{\cdots\}$$

Assume that φ, ϕ satisfy the conditions in the next theorem, then f has Lipschitz continuous gradient $\nabla f(x) = \nabla_x \phi(z(x), x)$.

Let φ be a lower semi-continuous convex function
$\phi(\cdot,\cdot) : \mathcal{Z} \times \mathcal{X} \to (-\infty, +\infty)$ is a convex function
$\phi(z,\cdot) : \Omega \to \Re$ is continuously differentiable on Ω for each z
$\nabla_x \phi(z,x)$ is continuous on $\mathrm{dom}(\varphi) \times \Omega$
For every $x' \in \Omega$, $z(x')$ is unique

Consider the convex function $f : \Omega \to [-\infty, +\infty)$ defined by

$$f(x) = \inf_{z \in \mathcal{Z}} \{\varphi(z) + \phi(z,x)\}, \quad x \in \Omega. \tag{14}$$

Condition: The minimizer $z(x)$ is unique for each x and is bounded on a compact set.

Theorem 18

(i) *If* $\exists$ *an open neighborhood* $\mathcal{N}_x$ *of* x *such that* $z(\cdot)$ *is continuous on* $\mathcal{N}_x$*, then* f *is continuously differentiable on* $\mathcal{N}_x$ *and*

$$\nabla f(x') = \nabla_x \phi(z(x'), x'), \quad \forall\, x' \in \mathcal{N}_x.$$

(ii) *Suppose that* $z(\cdot)$ *is bounded on any nonempty compact subset of* $\mathcal{N}$*. Assume that for any* $z \in \mathrm{dom}(\varphi)$*,* $\nabla_x \phi(z, \cdot)$ *is Lipschitz continuous on* $\mathcal{N}$ *and* $\exists\, \Sigma \succeq 0$ *such that for all* $x \in \mathcal{N}$*,*

$$\Sigma \succeq \mathcal{H}, \quad \forall\, \mathcal{H} \in \partial^2_{xx} \phi(z, x).$$

Then, $\nabla f(\cdot)$ *is Lipschitz continuous on* $\mathcal{X}$ *with the Lipschitz constant* $||\Sigma\|_2$ *(the spectral norm of* Σ*) and for any* $x \in \mathcal{X}$*,*

$$\Sigma \succeq \mathcal{G}, \quad \forall\, \mathcal{G} \in \partial^2_{xx} f(x),$$

where $\partial^2_{xx} f(x)$ *denotes the generalized Hessian of* f *at* x*.*

$$\min\{p(x_1) + \varphi(z) + \phi(z, x) \mid z \in \mathcal{Z},\ x \in \mathcal{X}\}$$

Apply inexact APG to

$$\min\{F(x) := p(x_1) + f(x) \mid x \in \mathcal{X}\}.$$

where $f(x) = \min_z\{\varphi(z) + \phi(z, x)\}$. Since $\nabla f(\cdot)$ is Lipschitz cont., $\exists$ PSD linear operator $\mathcal{Q} : \mathcal{X} \to \mathcal{X}$ such that

$$\mathcal{Q} \ \succeq \ \mathcal{G}, \quad \forall\, \mathcal{G} \in \partial^2 f(x),\ \forall x \in \mathcal{X}.$$

and $\mathcal{Q}_{ii} \succ 0$ for all i. We have

$$f(x) \leq q_k(x) := f(\bar{x}^k) + \langle \nabla f(\bar{x}^k),\, x - \bar{x}^k \rangle + \frac{1}{2}\|x - \bar{x}^k\|^2_{\mathcal{Q}}$$

$\nabla f(\bar{x}^k) = \nabla_x \phi(z(\bar{x}^k), \bar{x}^k)$.

Algorithm 2. Input $\bar{x}^1 = x^0 \in \mathrm{dom}\,(p) \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_s$, $t_1 = 1$. Let $\{\epsilon_k\}$ be a nonnegative summable sequence. Iterate

1. Compute via inexact sGS

$$z^k = z(\bar{x}^k) = \mathrm{argmin}_z \Big\{ \varphi(z) + \phi(z, \bar{x}^k) \Big\} \quad \text{(elimination via Danskin)}$$

$$x^k \approx \mathrm{argmin}_x \Big\{ p(x_1) + q_k(x) + \frac{1}{2}\|x - \bar{x}^k\|^2_{\mathcal{Q}\mathrm{SGS}} - \langle \Delta(\widehat{\delta}^k, \delta^k),\, x \rangle \Big\}$$

where the error vectors $\widehat{\delta}^k, \delta^k$ incurred in the forward and backward GS sweeps satisfy

$$\max\{\|\widehat{\delta}^k\|, \|\delta\|^k\} \leq \frac{\epsilon_k}{\sqrt{2}t_k}$$

2. Compute $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$, $\bar{x}^{k+1} = x^k + \left(\frac{t_k - 1}{t_{k+1}}\right)(x^k - x^{k-1})$.

Theorem 19

Let $\mathcal{H} = \mathcal{Q} + \mathcal{Q}^{SGS}$ *and* $\beta = 2\|\mathcal{D}^{-1/2}\| + \|\mathcal{H}^{-1/2}\|$. *The sequence* $\{(z^k, x^k)\}$ *generated by Algorithm 2 satisfies*

$$0 \le F(x^k) - F(x^*) \le \frac{4}{(k+1)^2}\left(\sqrt{\tau} + \beta\bar{\epsilon}_k\right)^2$$

where $\tau = \frac{1}{2}\|x^0 - x^*\|_{\mathcal{H}}^2$, $\bar{\epsilon}_k = \sum_{j=1}^k \epsilon_j$.

For the dual problem: $\varphi(Z) = \delta^*_{\mathcal{P}}(-Z),\ p(S) = \delta^*_{\mathcal{S}^n_+}(-S)$,

$$\phi(Z; S, y) = \frac{1}{2}\|Z + S + \mathcal{A}^* y + G\|^2 - \langle b,\ y\rangle$$

$$Z^k := \text{argmin}_Z\{\varphi(Z) + \phi(Z; (\bar{S}, \bar{y})^k)\} = -\mathrm{Prox}_{\delta^*_{\mathcal{P}}}(\bar{S}^k + \mathcal{A}^*\bar{y}^k + G)$$

$$= \Pi_{\mathcal{P}}(\bar{S}^k + \mathcal{A}^*\bar{y}^k + G) - (\bar{S}^k + \mathcal{A}^*\bar{y}^k + G)$$

$$\nabla f((\bar{S}, \bar{y})^k) = \nabla_{(S,y)}\phi(Z^k; (\bar{S}, \bar{y})^k) = \begin{bmatrix} Z^k + \bar{S}^k + \mathcal{A}^*\bar{y}^k + G \\ \mathcal{A}(Z^k + \bar{S}^k + \mathcal{A}^*\bar{y}^k + G) \end{bmatrix}$$

$$\mathcal{Q} = \begin{bmatrix} I & \mathcal{A}^* \\ \mathcal{A} & \mathcal{A}\mathcal{A}^* \end{bmatrix}$$

$$q_k(S, y) = \frac{1}{2}\|Z^k + S + \mathcal{A}^* y + G\|^2$$

Step 1.

(Danskin) $Z^k = \text{argmin}_Z\{\boldsymbol{F}(\boldsymbol{Z}, \bar{\boldsymbol{S}}^{\boldsymbol{k}}, \bar{\boldsymbol{y}}^{\boldsymbol{k}})\}$ (Projection onto $\mathcal{P}$)

(SGS)
$$\begin{cases} \widehat{y}_E^k = \text{argmin}_y\{\boldsymbol{F}(\boldsymbol{Z}^{\boldsymbol{k}}, \boldsymbol{S}^{\boldsymbol{k}}, \boldsymbol{y}) - \langle \widehat{\delta}^k, y\rangle\} & \text{(Chol or CG)} \\ S^k = \text{argmin}_S\{\boldsymbol{F}(\boldsymbol{Z}^{\boldsymbol{k}}, \boldsymbol{S}, \widehat{\boldsymbol{y}}^{\boldsymbol{k}})\} & \text{(Projection onto } \mathbb{S}^n_+) \\ y^k = \text{argmin}_y\{\boldsymbol{F}(\boldsymbol{Z}^{\boldsymbol{k}}, \boldsymbol{S}^{\boldsymbol{k}}, \boldsymbol{y}) - \langle \delta^k, y\rangle\} & \text{(Chol or CG)} \end{cases}$$

Step 2. Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $\beta_k = \frac{t_k - 1}{t_{k+1}}$. Compute

$$(\bar{S}, \bar{y})^{k+1} = (1+\beta_k)(S, y)^k - \beta_k (S, y)^{k-1}$$

- We compare the performance of our ABCD against the following methods for solving LSSDP:
  - BCD, APG
  - eARBCG (an enhanced accelerated randomized block coordinate gradient method) [Lin, Lu, Xiao, SIOPT 2015]
- We test the algorithms on LSSDP problem (**P**) by taking $G = -C$ for the data arising from various classes of SDP of the form (**SDP**).

SDP problem sets

Let $\mathcal{P} = \{X \in \mathcal{S}^n \mid X \geq 0\}$.

- SDP relaxation of a binary integer nonconvex quadratic (BIQ) programming:

$$
\begin{aligned}
\min \quad & \tfrac{1}{2}\langle Q,\, Y\rangle + \langle c,\, x\rangle \\
\text{s.t.} \quad & \operatorname{diag}(Y) - x = 0, \quad \alpha = 1, \\
& X = \begin{bmatrix} Y & x \\ x^T & \alpha \end{bmatrix} \in \mathbb{S}^n_+, \quad X \in \mathcal{P}
\end{aligned}
$$

- SDP relaxation $\theta_+(G)$ of the maximum stable set problem of a graph G with edge set $\mathcal{E}$:

$$
\max\{\langle ee^T,\, X\rangle \mid X_{ij} = 0, (i,j) \in \mathcal{E},\, \langle I,\, X\rangle = 1,\, X \in \mathbb{S}^n_+,\, X \in \mathcal{P}\}
$$

- SDP relaxation of clustering problems (RCPs):

$$
\min \left\{\langle W,\, I - X\rangle \mid Xe = e, \langle I,\, X\rangle = K, X \in \mathbb{S}^n_+, X \in \mathcal{P}\right\}
$$

SDP problem sets

- SDP arising from computing lower bounds for quadratic assignment problems (QAPs):

$$
\begin{aligned}
v := \min \quad & \langle B \otimes A,\, Y \rangle \\
\text{s.t.} \quad & \textstyle\sum_{i=1}^{n} Y^{ii} = I, \quad \langle I,\, Y^{ij} \rangle = \delta_{ij} \quad \forall\, 1 \le i \le j \le n, \\
& \langle E,\, Y^{ij} \rangle = 1 \quad \forall\, 1 \le i \le j \le n, \\
& Y \in \mathbb{S}_{+}^{n^2},\ Y \in \mathcal{P}
\end{aligned}
$$

where $\mathcal{P} = \{X \in \mathcal{S}^{n^2} \mid X \ge 0\}$.

- SDP relaxation of frequency assignment problems (FAPs):

Stop the algorithms after 25000 iterations, or

$$\eta = \max\{\eta_1, \eta_2\} < 10^{-6},$$

where $\eta_1 = \frac{\|b - \mathcal{A}X\|}{1+\|b\|}$, $\eta_2 = \frac{\|X - Y\|}{1+\|X\|}$

$X = \Pi_{\mathbb{S}^n_+}(\mathcal{A}^* y + Z + G)$, $Y = \Pi_{\mathcal{P}}(\mathcal{A}^* y + S + G)$

problem set (No.) \ solver	ABCD	APG	eARBCG	BCD
θ_+ (64)	64	64	64	11
FAP (7)	7	7	7	7
QAP (95)	95	95	24	0
BIQ (165)	165	165	165	65
RCP (120)	120	120	120	108
exBIQ (165)	165	141	165	10
Total (616)	616	592	545	201

Problem	m_E, m_I n	η ABCD | APG | eARBCG	time (hour:minute) ABCD | APG | eARBCG
1tc.2048	$18945, 0$ 2048	9.8-7 | 9.8-7 | 9.4-7	7:35 | 22:18 | 31:38
fap25	$2118, 0$ 2118	9.2-7 | 8.1-7 | 9.0-7	0:03 | 0:11 | 0:13
nug30	$1393, 0$ 900	9.6-7 | 9.9-7 | 1.4-6	0:10 | 1:12 | 7:21
tho30	$1393, 0$ 900	9.9-7 | 9.9-7 | 1.6-6	0:13 | 1:17 | 3:51
ex-gka5f	$501, 0.37M$ 501	9.8-7 | 1.6-6 | 9.9-7	0:24 | 2:26 | 4:00

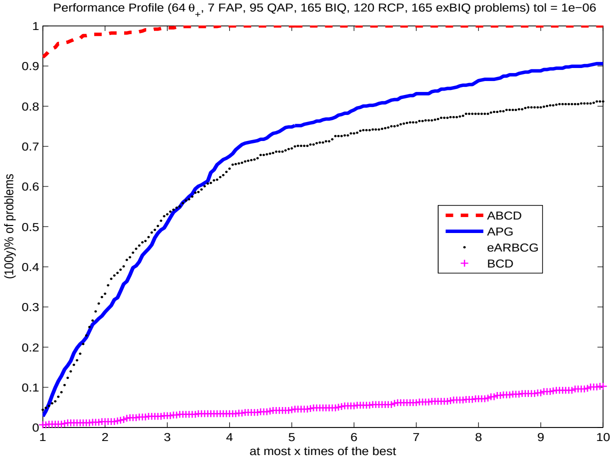


Figure: Performance profiles of ABCD, APG, eARBCG and BCD on $[1, 10]$

Proximal-point and augmented Lagrangian methods

- M. Fazel, T.K. Pong, D.F. Sun, and P. Tseng, *Hankel matrix rank minimization with applications to system identification and realization*, SIAM J. on Matrix Analysis and Applications, 34 (2013) 946–977.
- L. Chen, D.F. Sun, and K.C. Toh, *An efficient inexact symmetric Gauss-Seidel based majorized ADMM for high-dimensional convex composite conic programming*, Math. Prog., in print.
- R. T. Rockafellar. *Convex Analysis*. Princeton New Jersey, 1970.
- D.F. Sun, K.C. Toh and L.Q. Yang, *A convergent 3-block semi-proximal alternating direction method of multipliers for conic programming with 4-type constraints*, SIAM J. Optimization, 25 (2015) 882–915.
- Xinyuan Zhao, Defeng Sun, and K. C. Toh, *A Newton-CG augmented Lagrangian method for semidefinite programming*, SIAM J. on Optimization, 20 (2010) 1737–1765.

Most applications fit the general convex composite model:

$$\begin{array}{ll} \min\limits_{x\in\mathcal{X},y\in\mathcal{Y}} & \{\theta(x_1) + f(x_1,\ldots,x_m) + \phi(y_1) + g(y_1,\ldots,y_n)\} \\ \text{s.t.} & \sum_{i=1}^m \mathcal{A}_i^* x_i + \sum_{j=1}^n \mathcal{B}_j^* y_j = c \quad \rightarrow \; \mathcal{A}^* x + \mathcal{B}^* y = c \end{array} \tag{15}$$

$x = (x_1,\ldots,x_m) \in \mathcal{X} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_m$

$y = (y_1,\ldots,y_n) \in \mathcal{Y} := \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_n$

$\theta : \mathcal{X}_1 \rightarrow (-\infty,\infty]$, $\phi : \mathcal{Y}_1 \rightarrow (-\infty,\infty]$ are proper closed convex

$f : \mathcal{X} \rightarrow \Re$, $g : \mathcal{Y} \rightarrow \Re$ are convex Lip. cont. functions

$\mathcal{A}_i : \mathcal{Z} \rightarrow \mathcal{X}_i$, $\mathcal{B}_j : \mathcal{Z} \rightarrow \mathcal{Y}_i$ are given linear maps.

(GCCP) $\min\limits_{x\in\mathcal{X},y\in\mathcal{Y}} \left\{\theta(x_1) + f(x) + \phi(y_1) + g(y) \mid \mathcal{A}^* x + \mathcal{B}^* y = c\right\}$

Augmented Lagrangian function: Let $\sigma \in (0, +\infty)$.

$$\mathcal{L}_\sigma(x, y; z) := \theta(x_1) + f(x) + \phi(y_1) + g(y) + \langle z, \mathcal{A}^* x + \mathcal{B}^* y - c \rangle + \frac{\sigma}{2}\|\mathcal{A}^* x + \mathcal{B}^* y - c\|^2.$$

Augmented Lagrangian method (ALM):

Given an initial point $z^0 \in \mathcal{Z}$, perform the following iterations:

$$\begin{aligned} (x^{k+1}, y^{k+1}) &= \text{argmin}\, \mathcal{L}_\sigma(x, y; z^k), \\ z^{k+1} &= z^k + \tau\sigma(\mathcal{A}^* x^{k+1} + \mathcal{B}^* y^{k+1} - c) \end{aligned} \tag{16}$$

where $\tau \in (0, 2)$ is the steplength.

But difficult and expensive to solve the inner subproblems exactly or to high accuracy, especially in high-dimensional settings.

Sometimes it is easier to solve for x and y alternately, which motivates the use of ADMM. But they can still be difficult to solve.

Given cpc functions p, q. Consider

$$\text{(primal)} \quad \min\{p(x) + q(y) \mid \mathcal{A}^*x + \mathcal{B}^*y = c\}$$
$$\text{(dual)} \quad -\min\{\langle c,\, z\rangle + p^*(-\mathcal{A}z) + q^*(-\mathcal{B}z)\}$$

$$\mathcal{L}_\sigma(x, y; z) = p(x) + q(y) + \langle z,\, \mathcal{A}^*x + \mathcal{B}^*y - c\rangle + \frac{\sigma}{2}\|\mathcal{A}^*x + \mathcal{B}^*y - c\|^2$$

Pick $\mathcal{S} \succeq 0$, $\mathcal{T} \succeq 0$ and $\tau \in (0, \frac{1+\sqrt{5}}{2})$. Iterate

1. Comp. $x^{k+1} = \text{argmin}\{\mathcal{L}_\sigma(x, y^k; z^k) + \frac{1}{2}\|x - x^k\|^2_{\mathcal{S}}\}$
2. Comp. $y^{k+1} = \text{argmin}\{\mathcal{L}_\sigma(x^{k+1}, y; z^k) + \frac{1}{2}\|y - y^k\|^2_{\mathcal{T}}\}$
3. Comp. $z^{k+1} = z^k + \tau\sigma(\mathcal{A}^*x + \mathcal{B}^*y - c)$

Classical version: $\mathcal{S} = 0$, $\mathcal{T} = 0$ [Glowinski,Fortin,Marrocco,Gabay,Mercier]. Gabay showed that classical ADMM (for $\tau = 1$) is a special case of Douglas-Rachford splitting.

Eckstein and Bertsekas proved that DR splitting is an application of PPA on the dual via a specially-constructed splitting operator.

Convergence of generic 2-block semi-proximal ADMM (sPADMM)

By the monotonicity of ∂p, $\exists$ PSD operator Σ_p s.t.

$$\langle \xi - \xi', x - x' \rangle \geq \|x - x'\|^2_{\Sigma_p} \quad \forall\, \xi \in \partial p(x),\ \xi' \in \partial p(x')$$

Let Σ_q be similarly obtained.

Theorem 20 (FPST, SIMAX 2013)

Assume that constr. qualification holds and

$$\Sigma_p + \mathcal{S} + \sigma \mathcal{A}\mathcal{A}^* \succ 0, \quad \Sigma_q + \mathcal{T} + \sigma \mathcal{B}\mathcal{B}^* \succ 0$$

Then the seq. $\{(x^k, y^k)\}$ converges to an optimal sol. of (primal), and $\{z^k\}$ converges to an optimal sol. of (dual).

If y-part is absent, the convergence holds under $\tau \in (0, 2)$.

Proof is motivated by that of Glowinski, and Fortin & Glowinski for classical ADMM where $\mathcal{S} = 0, \mathcal{T} = 0$.

On designing convergent multi-block ADMM-type methods without restrictive assumptions.

- Direct extension of 2-block ADMM to multi-block setting is not guaranteed to converge [Chen,He,Ye,Yuan]. Convergent variant with Gaussian back substitution was designed in [He,Tao,Yuan]
- ADMM3c [Sun, T., Yang]: only requires an inexpensive extra step per iteration but is theoretically convergent and practically even faster.
- SCB-sPADMM and sGS-sPADMM [X.D.Li, Sun, T.]: can handle (15) where f and g are quadratic.
- Majorized iPADMM [M.Li, Sun, T.]: f and g are not necessarily quadratic functions; indefinite proximal terms.

Goal: to design inexact ADMM-type methods for solving the GCCP (15) in the multi-block and high-dimensional setting.

Majorized augmented Lagrangian function

Can consider f and g which are convex differentiable fun. with Lipschitz cont. gradients.

There exist self-adjoint PSD operators with $\widehat{\Sigma}_f \succeq \Sigma_f$, $\widehat{\Sigma}_g \succeq \Sigma_g$ such that for $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$,

$$f(x) \geq f(x') + \langle \nabla f(x'), x - x' \rangle + \tfrac{1}{2}\|x - x'\|^2_{\Sigma_f},$$

$$f(x) \leq \widehat{f}(x; x') := f(x') + \langle \nabla f(x'), x - x' \rangle + \tfrac{1}{2}\|x - x'\|^2_{\widehat{\Sigma}_f},$$

Similarly for $g(y)$ and $\widehat{g}(y; y')$.

At a given (x^k, y^k, z^k), consider the majorized augmented Lagrangian function defined by:

$$\mathcal{L}_\sigma(x, y; z^k) \;\leq\; \widehat{\mathcal{L}}_\sigma(x, y; z^k) := \theta(x_1) + \widehat{f}(x; x^k) + \phi(y_1) + \widehat{g}(y; y^k)$$
$$+\langle z^k, \mathcal{A}^* x + \mathcal{B}^* y - c \rangle + \tfrac{\sigma}{2}\|\mathcal{A}^* x + \mathcal{B}^* y - c\|^2$$

If f, g are quadratic fun., by taking $\widehat{\Sigma}_f = \Sigma_f$, $\widehat{\Sigma}_g = \Sigma_g$, and there is no majorization.

Choose self-adjoint PSD linear operations $\mathcal{S} : \mathcal{X} \to \mathcal{X}$, $\mathcal{T} : \mathcal{Y} \to \mathcal{Y}$ such that

$$\mathcal{N}_f := \widehat{\Sigma}_f + \mathcal{S} + \sigma \mathcal{A}\mathcal{A}^* \succ 0, \qquad \mathcal{N}_g := \widehat{\Sigma}_g + \mathcal{T} + \sigma \mathcal{B}\mathcal{B}^* \succ 0.$$

Suppose $\{(x^k, y^k, z^k)\}$ is a sequence in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Define

$$F_k(x) := \theta(x_1) + \frac{1}{2}\langle x, \mathcal{N}_f x\rangle - \langle l_f^k, x\rangle = \widehat{\mathcal{L}}_\sigma(x, y^k; z^k) + \frac{1}{2}\|x - x^k\|_{\mathcal{S}}^2 + \text{const}$$

$$G_k(y) := \phi(y_1) + \frac{1}{2}\langle y, \mathcal{N}_g y\rangle - \langle l_g^k, y\rangle = \widehat{\mathcal{L}}_\sigma(x^{k+1}, y; z^k) + \frac{1}{2}\|y - y^k\|_{\mathcal{T}}^2 + \text{const}$$

where

$$\begin{cases} -l_f^k := \nabla f(x^k) + \mathcal{A}z^k - \mathcal{N}_f x^k + \sigma \mathcal{A}(\mathcal{A}^* x^k + \mathcal{B}^* y^k - c) \\ -l_g^k := \nabla g(y^k) + \mathcal{B}z^k - \mathcal{N}_g y^k + \sigma \mathcal{B}(\mathcal{A}^* x^{k+1} + \mathcal{B}^* y^k - c). \end{cases}$$

Algorithm. Choose (x^0, y^0, z^0). Let $\{\varepsilon_k\}$ be a nonnegative summable seq. Perform the following steps at kth iteration.

1. Compute x^{k+1} and d_x^k such that

$$x^{k+1} \approx \bar{x}^{k+1} := \text{argmin}_x \Big\{ F_k(x) := \widehat{\mathcal{L}}_\sigma(x, y^k; z^k) + \frac{1}{2}\|x - x^k\|_{\mathcal{S}}^2 \Big\}$$

$$d_x^k \in \partial F_k(x^{k+1}) \quad \text{with } \|\mathcal{N}_f^{-\frac{1}{2}} d_x^k\| \leq \varepsilon_k. \tag{17}$$

2. Compute y^{k+1} and d_y^k such that

$$\begin{aligned} y^{k+1} \approx \bar{y}^{k+1} &:= \text{argmin}_y \Big\{ \widehat{\mathcal{L}}_\sigma(\bar{x}^{k+1}, y; z^k) + \frac{1}{2}\|y - y^k\|_{\mathcal{T}}^2 \Big\} \\ &= \text{argmin}_y \big\{ G_k(y) + \langle \sigma \mathcal{B}\mathcal{A}^*(\bar{x}^{k+1} - x^{k+1}),\ y \rangle \big\} \end{aligned}$$

$$d_y^k \in \partial G_k(y^{k+1}) \quad \text{with } \|\mathcal{N}_g^{-\frac{1}{2}} d_y^k\| \leq \varepsilon_k. \tag{18}$$

3. Compute $z^{k+1} = z^k + \tau\sigma(\mathcal{A}^* x^{k+1} + \mathcal{B}^* y^{k+1} - c)$ with fixed steplength $\tau \in (0, (1+\sqrt{5})/2)$

An inexact majorized 2-block sPADMM (imsPADMM)

Proposition 21

[Bound difference between (x^{k+1}, y^{k+1}) and $(\bar{x}^{k+1}, \bar{y}^{k+1})$ in terms of error tolerance]
Let $\{(x^k, y^k, z^k)\}$ be the seq. generated by imsPADMM, and $\{\bar{x}^k\}$, $\{\bar{y}^k\}$ be defined by (17) *and* (18). *For any $k \geq 0$, we have*

$$\|x^{k+1} - \bar{x}^{k+1}\|_{\mathcal{N}_f} \leq \varepsilon_k, \quad \|y^{k+1} - \bar{y}^{k+1}\|_{\mathcal{N}_g} \leq \varrho\varepsilon_k,$$

where $\varrho := 1 + \sigma\|\mathcal{N}_g^{-\frac{1}{2}}\mathcal{B}\mathcal{A}^*\mathcal{N}_f^{-\frac{1}{2}}\|$.

Theorem 22

Suppose f and g have Lipschitz cont. gradients, and

$$\Sigma_f + \mathcal{S} + \sigma\mathcal{A}\mathcal{A}^* \succ 0, \quad \Sigma_g + \mathcal{T} + \sigma\mathcal{B}\mathcal{B}^* \succ 0.$$

Suppose constraint qualification holds, i.e., $\exists$ $(x, y) \in$ ri (dom (p) $\times$ dom (q)) s.t. $\mathcal{A}^*x + \mathcal{B}^*y = c$. Then $\{(x^k, y^k)\}$ converges to an optimal sol. of (15) *and $\{z^k\}$ converges to an optimal sol. of the dual.*

Algorithm 2. Choose step-length $\tau \in (0, (1+\sqrt{5})/2)$, $\{\tilde{\varepsilon}_k\}$ a nonnegative summable seq. Perform the following steps.

[Step 1a] (*Backward GS sweep*) Compute for $i = m, \ldots, 2$,

$$\tilde{x}_i^{k+1} \approx \arg\min_{x_i \in \mathcal{X}_i} \left\{ \widehat{\mathcal{L}}_\sigma(x_{\le i-1}^k, x_i, \tilde{x}_{\ge i+1}^{k+1}, y^k) + \frac{1}{2}\|x_i - x_i^k\|_{S_i}^2 \right\}$$

$$\tilde{\delta}_i^k \in \partial_{x_i} \widehat{\mathcal{L}}_\sigma(x_{\le i-1}^k, \tilde{x}_i^{k+1}, \tilde{x}_{\ge i+1}^{k+1}, y^k) + S_i(\tilde{x}_i^{k+1} - x_i^k) \text{ with } \|\tilde{\delta}_i^k\| \le \tilde{\varepsilon}_k.$$

[Step 1b] (*Forward GS sweep*) Compute for $i = 1, \ldots, m$,

$$x_i^{k+1} \approx \arg\min_{x_i \in \mathcal{X}_i} \left\{ \widehat{\mathcal{L}}_\sigma(x_{\le i-1}^{k+1}, x_i, \tilde{x}_{\ge i+1}, y^k) + \frac{1}{2}\|x_i - x_i^k\|_{S_i}^2 \right\}$$

$$\delta_i^k \in \partial_{x_i} \widehat{\mathcal{L}}_\sigma(x_{\le i-1}^{k+1}, x_i^{k+1}, \tilde{x}_{\ge i+1}, y^k) + S_i(x_i^{k+1} - x_i^k) \text{ with } \|\delta_i^k\| \le \tilde{\varepsilon}_k.$$

Algorithm 2 (Continued).

[Step 2a] (*Backward GS sweep*) for $j = n, \ldots, 2$,

$$\tilde{y}_j^{k+1} \approx \arg\min_{y_i \in \mathcal{Y}_i} \left\{ \widehat{\mathcal{L}}_\sigma(x^{k+1}, y_{\le j-1}^k, y_j, \tilde{y}_{\ge j+1}^{k+1}) + \frac{1}{2}\|y_j - y_j^k\|_{T_j}^2 \right\}$$

$\tilde{\gamma}_j^k \in \partial_{y_j} \widehat{\mathcal{L}}_\sigma(x^{k+1}, y_{\le j-1}^k, \tilde{y}_j^{k+1}, \tilde{y}_{\ge j+1}^{k+1}) + T_j(\tilde{y}_j^{k+1} - y_i^k)$ with $\|\tilde{\gamma}_j^k\| \le \tilde{\varepsilon}_k$.

[Step 2b] (*Forward GS sweep*) Compute for $j = 1, \ldots, n$,

$$y_j^{k+1} \approx \arg\min_{y_i \in \mathcal{Y}_i} \left\{ \widehat{\mathcal{L}}_\sigma(x^{k+1}, y_{\le j-1}^{k+1}, y_j, \tilde{y}_{\ge j+1}^{k+1}) + \frac{1}{2}\|y_j - y_j^k\|_{T_j}^2 \right\}$$

$\gamma_j^k \in \partial_{y_j} \widehat{\mathcal{L}}_\sigma(x^{k+1}, y_{\le j-1}^{k+1}, y_j^{k+1}, \tilde{y}_{\ge j+1}^{k+1}) + T_j(y_j^{k+1} - y_i^k)$ with $\|\gamma_j^k\| \le \tilde{\varepsilon}_k$.

[Step 3] Compute

$$z^{k+1} = z^k + \tau\sigma(\mathcal{A}^* x^{k+1} + \mathcal{B}^* y^{k+1} - c).$$

sGS-imsPADMM for solving GCCP (15)

Denote $\tilde{\delta}^k = (\tilde{\delta}_1^k, \ldots, \tilde{\delta}_m^k)$, $\delta^k = (\delta_1^k, \ldots, \delta_m^k)$. Similarly for $\tilde{\gamma}^k$, γ^k.

Proposition 23

Let $\mathcal{M}_f = \widehat{\Sigma}_f + \sigma\mathcal{A}\mathcal{A}^* + S$, $\mathcal{M}_g = \widehat{\Sigma}_g + \sigma\mathcal{B}\mathcal{B}^* + T$. *Then*

$$\mathcal{N}_f := \mathcal{M}_f + \mathcal{M}_f^{\mathsf{SGS}} \succ 0, \quad \mathcal{N}_g := \mathcal{M}_g + \mathcal{M}_g^{\mathsf{SGS}} \succ 0$$

The sequence $\{(x^k, y^k, z^k)\}$ *generated by sGS-imsPADMM satisfies*

$$d_x^k \in \partial_x \Big\{ \widehat{\mathcal{L}}_\sigma(x^{k+1}, y^k) + \tfrac{1}{2}\|x^{k+1} - x^k\|^2_{S + \mathcal{M}_f^{\mathsf{SGS}}} \Big\}$$
$$d_y^k \in \partial_y \Big\{ \widehat{\mathcal{L}}_\sigma(x^{k+1}, y^{k+1}) + \tfrac{1}{2}\|y^{k+1} - y^k\|^2_{T + \mathcal{M}_g^{\mathsf{SGS}}} \Big\}$$

with $\|\mathcal{N}_f^{-\frac{1}{2}} d_x^k\| \le \kappa \tilde{\varepsilon}_k, \quad \|\mathcal{N}_g^{-\frac{1}{2}} d_y^k\| \le \kappa' \tilde{\varepsilon}_k.$

where the const. κ depends on m and $\mathcal{M}_f$ (similarly for κ')

$d_x^k := \delta^k + \mathcal{M}_f^u(\mathcal{M}_f^d)^{-1}(\delta^k - \tilde{\delta}^k)$, $d_y^k := \gamma^k + \mathcal{M}_g^u(\mathcal{N}_g^d)^{-1}(\gamma^k - \tilde{\gamma}^k)$

The seq. $\{(x^k, y^k, z^k)\}$ generated by sGS-imsPADMM can be viewed as a seq. generated by imsPADMM with specifically chosen semi-proximal terms:

$$\mathcal{S} := S + \mathcal{M}_f^{\mathsf{SGS}}, \quad \mathcal{T} := T + \mathcal{M}_g^{\mathsf{SGS}}.$$

In particular, it always satisfies the admissible cond. (17) and (18) if we choose the summable seq. $\{\varepsilon_k\}$ such that $\varepsilon_k := \max(\kappa, \kappa')\tilde{\varepsilon}_k$.

- sGS-imsPADMM: an implementable approach to handle high-dimensional convex composite conic optimization problems.
- imsPADMM: the compact formulation can facilitate the convergence analysis of the sGS-imsPADMM.
- Cost saving can be done as before for the forward GS sweeps.

ADMM+ = our sGS-sPADMM applied to **dual DNN SDP problem:**
$\min\{-\langle b,\, y\rangle \mid \mathcal{A}^* y + S + Z = C, S \in \mathbb{S}^n_+, Z \in \mathbb{N}^n\}$

ADMM2 = sGS-sPADMM applied to reformulated dual prob:
$\min\{-\langle b,\, y\rangle \mid -\mathcal{A}^* y + U + S = C, U - Z = 0, S \in \mathbb{S}^n_+, Z \in \mathbb{N}^n\}$

ADMM3g = 3-block ADMM with Gaussian back substitution [He et al] applied to dual

sPADMM2-prim = sPADMM applied to reformulated primal prob:
$\min\{\langle C,\, X\rangle \mid \mathcal{A}X = b, X - V = 0, X \in \mathbb{S}^n_+, V \in \mathbb{N}^n\}$.

ADMM-prim = classical ADMM applied to reformulated primal prob:
$\min\{\langle C,\, X\rangle \mid \mathcal{A}X = b, X - U = 0, X - V = 0, U \in \mathbb{S}^n_+, V \in \mathbb{N}^n\}$.

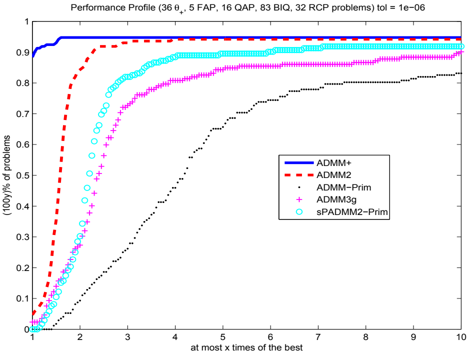


Figure: Performance profiles of various ADMM-type methods

Application of inexact SGS-ADMM in distance weighted discrimination (DWD)

n training samples: (x_i, y_i), $i = 1, \ldots, n$
$x_i \in \mathbb{R}^d$: d-dimensional feature vector
$y_i \in \{+1, -1\}$: corresponding class label.

In linear discrimination: separate the vectors in the two classes by a hyperplane $H = \{x \in \mathbb{R}^d \mid w^T x + \beta = 0\}$, where $w \in \mathbb{R}^d$ is the unit normal and $|\beta|$ is its distance to the origin.

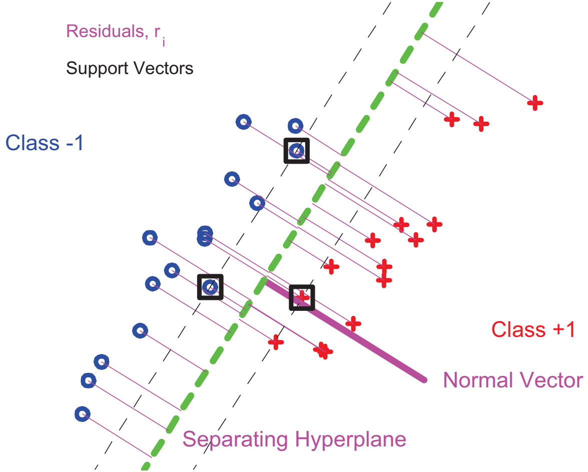
For binary classification:

$$y_i(\beta + x_i^T w) \geq 1 - \xi_i \quad \forall\, i = 1, ..., n,$$

where $\xi \geq 0$ is a slack variable to allow the possibility that the positive and negative data points may not be separated cleanly by the hyperplane.
In matrix-vector notation:

$$r \;:=\; Z^T w + \beta y + \xi \;\geq\; \mathbf{1}$$

where $Z = X \operatorname{diag} y \in \mathbb{R}^{d \times n}$ and $\mathbf{1} \in \mathbb{R}^n$ is the vector of ones.



Residuals, r_i
Support Vectors
Class -1
Class +1
Normal Vector
Separating Hyperplane

In SVM, w and β are chosen by maximizing the minimum residual:

$$\max\left\{\delta - C\langle \mathbf{1}, \xi\rangle \mid Z^T w + \beta y + \xi \geq \delta\mathbf{1},\ \xi \geq 0,\ w^T w \leq 1\right\}$$

where $C > 0$ is a parameter to control the penalty on ξ.

In DWD (Marron, Todd and Ahn, 2007), w and β are chosen by minimizing the sum of reciprocals of the r_i's:

$$\min\left\{ \sum_{i=1}^{n} \frac{1}{r_i} + C\langle \mathbf{1}, \xi\rangle \middle| \begin{array}{l} r = Z^T w + \beta y + \xi,\ r > 0,\ \xi \geq 0 \\ w^T w \leq 1,\ w \in \mathbb{R}^d \end{array} \right\}$$

It is proposed to avoid data piling at the supporting planes. The above DWD problem can be transformed to a second-order cone problem. Hence can be solved by interior-point methods.

$$\begin{aligned}\min \quad & \Phi(r,\xi) := \textstyle\sum_{i=1}^{n} \theta_q(r_i) + C\langle e,\, \xi\rangle \\ \text{s.t} \quad & Z^T w + \beta y + \xi - r = 0 \\ & \|w\| \leq 1,\ \xi \geq 0\end{aligned} \tag{19}$$

where $e \in \mathbb{R}^n$ is a given positive vector such that $\|e\|_\infty = 1$, and $e_i > 0$ is a nonuniform weight on the penalty term for each ξ_i.

$$\theta_q(t) = \begin{cases} \frac{1}{t^q} & \text{if } t > 0, \\ \infty & \text{if } t \leq 0. \end{cases}$$

$q > 0$ with most interested values $0.5, 1, 2, 4$.

Proposition 24

The dual of problem (19) *is given as follows:*

$$-\min_{\alpha}\left\{\Psi(\alpha) := \|Z\alpha\| - \kappa\sum_{i=1}^{n}\alpha_i^{\frac{q}{q+1}} \mid 0 \le \alpha \le Ce, \langle y,\, \alpha\rangle = 0\right\} \quad (20)$$

where $\kappa = \frac{q+1}{q} q^{\frac{1}{q+1}}$.

The feasible regions of (19) and (24) both have nonempty interiors. Thus optimal solutions for both problems exist and they satisfy the following optimality conditions:

$$Z^T w + \beta y + \xi - r = 0, \qquad \langle y,\, \alpha\rangle = 0,$$

$$r > 0,\ \alpha > 0, \quad \alpha \le Ce, \quad \xi \ge 0, \qquad \langle Ce - \alpha,\, \xi\rangle = 0,$$

$$\alpha_i = \frac{q}{r_i^{q+1}},\ i = 1, \ldots, n,$$

either $w = \frac{Z\alpha}{\|Z\alpha\|}$, or $Z\alpha = 0,\ \|w\|^2 \le 1.$

An equivalent formulation

Rewrite the primal model (19) as:

$$\begin{array}{ll} \min & \sum_{i=1}^{n} \theta_q(r_i) + C\langle e,\, \xi\rangle + \delta_B(w) + \delta_{\mathbb{R}^n_+}(\xi) \\ \text{s.t.} & Z^T w + \beta y + \xi - r = 0, \;\; w \in \mathbb{R}^d,\; r, \xi \in \mathbb{R}^n, \end{array}$$

where $B = \{w \in \mathbb{R}^d \mid \|w\| \leq 1\}$.

Introducing an auxiliary variable $u = w$:

$$\begin{array}{ll} \min & \sum_{i=1}^{n} \theta_q(r_i) + C\langle e,\, \xi\rangle + \delta_B(u) + \delta_{\mathbb{R}^n_+}(\xi) \\ \text{s.t.} & Z^T w + \beta y + \xi - r = 0 \\ & D(w - u) = 0, \;\; w, u \in \mathbb{R}^d,\; \beta \in \mathbb{R},\; r, \xi \in \mathbb{R}^n, \end{array} \tag{21}$$

where $D \in \mathbb{R}^{d\times d}$ is a given positive scalar multiple of the identity matrix which is introduced to scale the variables.

This is a linearly constrained nonsmooth convex programming problem with three blocks of variables: (w, β), r, (u, ξ).

Augmented Lagrangian function associated with (21):

$$L_\sigma(r, w, \beta, \xi, u; \alpha, \rho) = \sum_{i=1}^n \theta_q(r_i) + C\langle e, \xi\rangle + \delta_B(u) + \delta_{\mathbb{R}^n_+}(\xi)$$

$$+\frac{\sigma}{2}\| - r + Z^T w + \beta y + \xi - \sigma^{-1}\alpha\|^2 + \frac{\sigma}{2}\|D(w - u) - \sigma^{-1}\rho\|^2 - \frac{1}{2\sigma}\|\alpha\|^2 - \frac{1}{2\sigma}\|\rho\|^2.$$

Given an initial point, performance the following steps in each iteration. Treat (r, w, β) as first block, (u, ξ) as second block.

Step 1a. Compute

$$(\bar{w}^{k+1}, \bar{\beta}^{k+1}) \approx \min_{w,\beta} \left\{L_\sigma(r^k, w, \beta, \xi^k, u^k; \alpha^k, \rho^k)\right\}$$

Step 1b. Compute $r^{k+1} \approx \min_{r\in\mathbb{R}^n} \left\{L_\sigma(r, \bar{w}^{k+1}, \bar{\beta}^{k+1}, \xi^k, u^k; \alpha^k, \rho^k)\right\}$

Step 1c. Compute

$$(w^{k+1}, \beta^{k+1}) \approx \min_{w,\beta} \left\{L_\sigma(r^{k+1}, w, \beta, \xi^k, u^k; \alpha^k, \rho^k)\right\}$$

Note: the subproblems need not be solved exactly as long as they satisfy some prescribed accuracy.

Step 2. Compute

$$(u^{k+1}, \xi^{k+1}) = \min_{u,\xi} L_\sigma(r^{k+1}, w^{k+1}, \beta^{k+1}, \xi, u; \alpha^k, \rho^k)$$

$$u^{k+1} = \begin{cases} g^k & \text{if } \|g^k\| \leq 1,\ g^k = w^{k+1} - \sigma^{-1} D^{-1} \rho^k \\ g^k/\|g^k\| & \text{otherwise} \end{cases}$$

$$\xi^{k+1} = \Pi_{\mathbb{R}^n_+}\Big(r^{k+1} - Z^T w^{k+1} - y\beta^{k+1} + \sigma^{-1}\alpha^k - \sigma^{-1} Ce\Big)$$

Step 3. Compute

$$\begin{aligned} \alpha^{k+1} &= \alpha^k - \tau\sigma(Z^T w^{k+1} + y\beta^{k+1} + \xi^{k+1} - r^{k+1}) \\ \rho^{k+1} &= \rho^k - \tau\sigma D(w^{k+1} - u^{k+1}), \end{aligned}$$

where $\tau \in (0, (1+\sqrt{5})/2)$ is the steplength which is typically chosen to be 1.618.

Convergence result for inexact sGS-ADMM

Theorem 25 (Chen-Sun-Toh, 2015)

Suppose (21) *has at least one solution. Let* $\{(r^k, w^k, \beta^k, \xi^k, u^k; \alpha^k, \rho^k)\}$ *be the sequence generated by the inexact sGS-ADMM in Algorithm 1. Then* $\{(r^k, w^k, \beta^k, \xi^k, u^k)\}$ *converges to an optimal solution of* (21) *and* $\{(\alpha^k, \rho^k)\}$ *converges to an optimal solution of its dual.*

Numerical experiments: Comparison between solvers

exponent $q = 2$			sGS-ADMM			directADMM			IPM		
Data	n	d	Time (s)	Iter	Err (%)	Time (s)	Iter	Err	Time (s)	Iter	Err
covtype	581012	54	74.69	368	23.74	336.85	2343	23.74	-	-	-
gisette	6000	5000	48.04	312	0.00	273.49	5000	1.07	2508.89	55	0.00
ijcnn1	35000	22	2.95	233	7.94	8.28	1004	7.88	2002.77	38	7.98
mush.	8124	112	1.40	301	0.00	16.33	5000	0.10	650.16	52	0.00
real-sim	72309	20958	29.25	174	1.51	19.26	181	1.51	-	-	-
w8a	49749	300	8.41	543	1.13	52.77	5000	2.66	-	-	-
rcv1	20242	47236	20.61	121	0.83	25.44	267	0.78	7547.28	43	0.79
prostate	102	6033	0.11	41	0.00	0.41	301	0.00	8.78	28	0.00*
farm-ads	4143	54877	5.60	80	0.41	5.36	113	0.41	383.43	41	0.41*
dorothea	800	100000	6.41	253	0.00	80.14	5000	0.00	25.89	35	0.00
url-svm	256000	3231961	1198.81	384	0.50	1406.38	770	0.50	-	-	-

- Interior point method is always the slowest
- In terms of iterations, our inexact sGS-ADMM has the best performance
- For 7 data sets, the number of iterations required by the directly extended ADMM hits the maximum allowed, probably implying nonconvergence of the method.

Semismooth Newton based augmented Lagrangian method (SNAL) with applications to big sparse optimization

[X.Y. Zhao, D.F. Sun, and K.C. Toh] A Newton-CG augmented Lagrangian method for semidefinite programming, SIOPT, 2010

[L.Q. Yang, D.F. Sun, and K.C Toh] SDPNAL+: a majorized semismooth Newton-CG augmented Lagrangian method for semidefinite programming with nonnegative constraints, MPC, 2015

[X.D. Li, D.F. Sun, and K.C. Toh] QSDPNAL: A two-phase proximal augmented Lagrangian method for convex quadratic semidefinite programming, submitted, 2015

[X.D. Li, D.F. Sun, and K.C. Toh] QP-PAL: A 2-phase proximal augmented Lagrangian method for high dimensional convex quadratic and linear programming problems, technical report, 2016

[X.D. Li, D.F. Sun, and K.C. Toh] An efficient linearly convergent semismooth Netwon-CG augmented Lagrangian method for Lasso problems, submitted, 2016

[X.D. Li, D.F. Sun, and K.C. Toh] A semismooth Netwon augmented Lagrangian method for fused Lasso problems, technical report, 2016

Let $\mathcal{X}, \mathcal{Y}$ be real finite dim. real inner product spaces. Consider

$$\textbf{(P)} \quad \min \left\{ f(x) := h(\mathcal{A}x) + \theta(x) \right\}$$

$\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ is a linear map
$h : \mathcal{Y} \to (\infty, \infty]$, convex differentiable
$\theta : \mathcal{X} \to (-\infty, \infty]$, cpc fun.

The dual (ignoring the minus sign) is

$$\textbf{(D)} \quad \min_{\xi \in \mathcal{Y},\, u \in \mathcal{X}} \left\{ h^*(\xi) + \theta^*(u) \mid \mathcal{A}^*\xi + u = 0 \right\}$$

The KKT conditions are:

$$\mathcal{A}^*\xi + u = 0, \quad u \in \partial\theta(x), \quad \xi \in \partial h(Ax)$$

$$\Leftrightarrow \quad \mathcal{A}^*\xi + u = 0, \quad x = \mathsf{P}_\theta(x + u), \quad \xi = \nabla h(Ax)$$

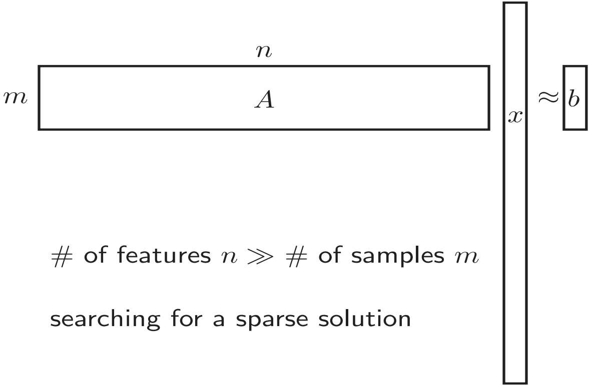
Examples of smooth loss function h:

- Linear regression: $\|y - b\|^2$
- Logistic regression: $\log(1 + \exp(-y^T b))$
- Support vector machine (SVM dual): $\frac{1}{2}\|y\|^2$

Examples of regularizer θ:

- Lasso: $\|x\|_1$
- Fused Lasso: $\|x\|_1 + \lambda \sum_{i=1}^{n-1} |x_i - x_{i+1}|$
- Clustered Lasso: $\|x\|_1 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{i-1} |x_i - x_j|$
- Ridge: $\|x\|_2^2$
- Elastic net: $\|x\|_1 + \lambda \|x\|_2^2$
- Group Lasso: $\sum_{i=1}^{N} \|x_{g_i}\|_2$
- Fused Group Lasso:
- SVM: $-\langle x, \mathbf{1}\rangle + \delta_B(x)$ where $B = \{x \mid \langle b, x\rangle = 0, 0 \leq x \leq \rho\}$.
- etc

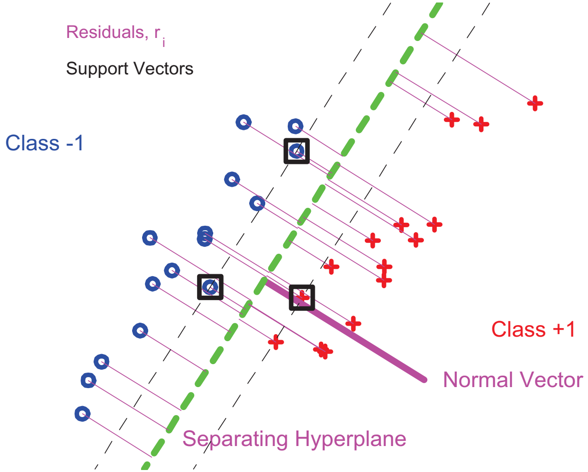
Sparse regression:



# of features $n \gg$ # of samples m

searching for a sparse solution

Example: Support vector machine



1. $h : \mathcal{Y} \to \Re$ has a $1/\alpha_h$-Lipschitz continuous gradient:

$$\|\nabla h(y_1) - \nabla h(y_2)\| \leq (1/\alpha_h)\|y_1 - y_2\| \quad \forall\, y_1, y_2 \in \mathcal{Y}$$

2. h is essentially locally strongly convex [Goebel & Rockafellar, 2008]: for any compact convex set $K \subset \operatorname{dom} \partial f$, $\exists\, \sigma_K > 0$ s.t. for all $\alpha, \beta \in [0, 1]$, $\alpha + \beta = 1$, $y_1, y_2 \in K$:

$$\alpha h(y_1) + \beta h(y_2) \geq h(\alpha y_1 + \beta y_2) + \frac{\sigma_K}{2}\alpha\beta\|y_1 - y_2\|^2$$

Under the above assumptions on h, we have

a. h^*: strongly convex with const. α_h and $\mathcal{D}_{h^*} := \operatorname{int}(\operatorname{dom} h^*) \neq \emptyset$

b. $\partial h^*(y) = \emptyset$ when $y \notin \mathcal{D}_{h^*}$

c. h^*: essentially smooth (diff. on $\mathcal{D}_{h^*}$ and $\|\nabla h^*(y_i)\| \uparrow \infty$ whenever $\{y_i\} \subset \mathcal{D}_{h^*} \to y \in \operatorname{bdry}(\mathcal{D}_{h^*})$)

d. ∇h^*: locally Lipschitz continuous on $\mathcal{D}_{h^*}$

Only need to focus on $\mathcal{D}_{h^*}$!

Proximal-point algorithm for (P)

Given $\sigma > 0$, the Moreau-Yosida regularization of f is given by

$$\mathbf{F}_\sigma(\widehat{x}) := \min\left\{ f(x) + \frac{1}{2\sigma}\|x - \widehat{x}\|^2 \right\}$$

Denote the unique minimizer by $\mathbf{P}_\sigma(\widehat{x})$. $\mathbf{F}_\sigma$ is differentiable and

$$\begin{aligned} \nabla \mathbf{F}_\beta(x) &= \frac{1}{\sigma}(x - \mathbf{P}_\sigma(x)) \\ \|\mathbf{P}_\sigma(x) - \mathbf{P}_\sigma(x')\| &\leq \|x - x'\| \quad \forall\, x, x' \\ \min f(x) &\Leftrightarrow \min \mathbf{F}_\sigma(x) \end{aligned}$$

Clearly $\min f \leq \min \mathbf{F}_\sigma$. Conversely, $\min \mathbf{F}_\sigma \leq \mathbf{F}_\sigma(\widehat{x}) \leq f(\widehat{x})$ for all $\widehat{x} \Rightarrow \min \mathbf{F}_\sigma \leq \min f$.

PPA is a gradient method to solve $\min \mathbf{F}_\sigma(x)$:

$$x^{k+1} \approx x^k - \sigma_k \nabla \mathbf{F}_{\sigma_k}(x^k) = \mathbf{P}_{\sigma_k}(x^k).$$

To compute $\mathbf{P}_\sigma(x^k)$, we solve

$$\mathbf{F}_\sigma(x^k) = \min\left\{f(x) + \frac{1}{2\sigma}\|x - x^k\|^2\right\}$$

$$= \frac{1}{2\sigma}\|x^k\|^2 - \min_{\xi}\left\{\begin{array}{l} h^*(\xi) + \frac{1}{2\sigma}\|x^k - \sigma\mathcal{A}^*\xi\|^2 \\ -\frac{1}{\sigma}\mathsf{M}_{\sigma\theta}(x^k - \sigma\mathcal{A}^*\xi) \end{array}\right\} \quad \left(\begin{array}{c}\text{by strong}\\ \text{duality}\end{array}\right)$$

$$= \frac{1}{2\sigma}\|x^k\|^2 - \boxed{\min_\xi \left\{h^*(\xi) + \sigma\mathsf{M}_{\theta^*/\sigma}(\sigma^{-1}x^k - \mathcal{A}^*\xi)\right\}}$$

for dual solution ξ^k. Then set $\mathbf{P}_\sigma(x^k) = \mathsf{P}_{\sigma\theta}(x^k - \sigma\mathcal{A}^*\xi^k)$.

Optimality condition is

$$0 \in \partial h^*(\xi) - \mathcal{A}\mathsf{P}_{\sigma\theta}(x^k - \sigma\mathcal{A}^*\xi).$$

The Lagrangian function for (**D**):

$$l(\xi, u; x) = h^*(\xi) + \theta^*(u) - \langle x,\, \mathcal{A}^*\xi + u\rangle, \quad \forall\, (\xi, u, x) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{X}.$$

Given $\sigma > 0$, the augmented Lagrangian function for (**D**):

$$\mathcal{L}_\sigma(\xi, u; x) = l(\xi, u; x) + \frac{\sigma}{2}\|\mathcal{A}^*\xi + u\|^2$$

The proximal mapping $\mathsf{P}_\theta(x)$:

$$\mathsf{P}_\theta(x) = \arg\min_{u\in\mathcal{X}} \left\{\theta(x) + \frac{1}{2}\|u - x\|^2\right\}.$$

Given $C \subset \mathcal{X}$, let $\mathsf{dist}(x, C) := \inf_{x'\in C} \|x - x'\|$ for any $x \in \mathcal{X}$.

An inexact augmented Lagrangian method of multipliers.

Given $\sum \varepsilon_k < \infty$, $\sigma_0 > 0$, choose $(\xi^0, u^0, x^0) \in \mathcal{D}_{h^*} \times \operatorname{dom} p^* \times \mathcal{X}$. Iterate

1. Compute

$$(\xi^{k+1}, u^{k+1}) \approx \operatorname{argmin}\Big\{\mathcal{L}^k(\xi, u) := \mathcal{L}_{\sigma_k}(\xi, u; x^k)\Big\}.$$

2. Compute $x^{k+1} = x^k - \sigma_k(\mathcal{A}^*\xi^{k+1} + u^{k+1})$;
   update $\sigma_{k+1} \uparrow \sigma_\infty \leq \infty$.

The stopping criterion for inner subproblem

$$\text{(A)} \quad \mathsf{dist}(0, \partial \mathcal{L}^k(\xi^{k+1}, u^{k+1})) \leq \varepsilon_k / \max(1, \sqrt{\sigma_k/\alpha_h})$$

Theorem 26 (Global convergence)

Suppose that the solution set to **(P)** *is nonempty. Then* $\{x^k\}$ *is bounded and converges to an optimal sol. of* **(P)**. *In addition,* $\{(\xi^k, u^k)\}$ *is also bounded and converges to the unique optimal sol.* $(\xi^*, u^*) \in \mathcal{D}_{h^*} \times \mathrm{dom}(p^*)$ *of* **(D)**.

Assumption 27 (Error bound)

For a maximal monotone operator $\mathcal{T} : \mathcal{X} \rightrightarrows \mathcal{Y}$ *with* $\mathcal{T}^{-1}(0) \neq \emptyset$, $\exists\, \varepsilon > 0$ *and* $a > 0$ *s.t.*

$$\mathrm{dist}(\xi, \mathcal{T}^{-1}(0)) \leq a\|\eta\| \quad \forall\, \eta \in B_\varepsilon \quad \text{and} \quad \forall\, \xi \in \mathcal{T}^{-1}(\eta)$$

where $B_\varepsilon = \{y \in \mathcal{Y} \mid \|y\| \leq \varepsilon\}$. *The constant* a *is called the error bound modulus associated with* $\mathcal{T}$.

It holds for the following cases:

1. $\mathcal{T}$ is a polyhedral multifunction [Robinson, 1981].
2. $\mathcal{T}_l$ and $\mathcal{T}_f$ of LASSO, fused LASSO and elastic net regularized LS problems (piecewise quadratic problems [J. Sun, 1986]).
3. $\mathcal{T}_f$ of ℓ_1 or elastic net regularized logistic regression [Luo and Tseng, 1992; Tseng and Yun, 2009].

Stopping criterion for the local convergence analysis

$$
\text{(B)} \qquad \text{dist}(0, \partial \mathcal{L}^k(\xi^{k+1}, u^{k+1})) \leq \frac{\delta_k}{\max(1, \sigma_k/\alpha_h)} \min\{1, \|x^{k+1} - x^k\|\}, \quad \textstyle\sum_{k=0}^{\infty} \delta_k < \infty
$$

Theorem 28

Assume that the solution set Ω to **(P)** *is nonempty and error bound assumption holds for $\mathcal{T}_f$ with modulus a_f. Then, $\{x^k\}$ is convergent and, for all k sufficiently large,*

$$
\text{dist}(x^{k+1}, \Omega) \;\leq\; \theta_k \text{dist}(x^k, \Omega)
$$

where $\theta_k = \Big(\frac{a_f}{(a_f^2 + \sigma_k^2)^{1/2}} + 2\delta_k\Big)/(1 - \delta_k) \to \theta_\infty = \frac{a_f}{(a_f^2 + \sigma_\infty^2)^{1/2}} < 1.$

Moreover, the conclusions of Theorem 26 about $\{(\xi^k, y^k)\}$ are valid.

$$\min_{\xi,u} \left\{ \mathcal{L}^k(\xi,u) := \mathcal{L}_\sigma(\xi,u;x^k) \right\} \;=\; \min_{\xi} \psi(\xi)$$

$$\psi(\xi) \;:=\; \inf_u \{\mathcal{L}_\sigma(\xi,u,x^k)\} \;=\; \boxed{h^*(\xi) + \sigma \mathsf{M}_{\theta^*/\sigma}(\sigma^{-1}x^k - \mathcal{A}^*\xi)}$$

which is exactly the dual of the PPA subproblem.

$\psi(\cdot)$: strongly convex and continuously differentiable on $\mathcal{D}_{h^*}$ with

$$\nabla\psi(\xi) = \nabla h^*(\xi) - \mathcal{A}\,\mathsf{P}_{\sigma\theta}(x^k - \sigma\mathcal{A}^*\xi) \quad \forall\, \xi \in \mathcal{D}_{h^*}.$$

Solving nonsmooth equation:

$$\nabla\psi(\xi) = 0 \quad \xi \in \mathcal{D}_{h^*}$$

Lipschitz continuous mapping: ∇h^*, $\mathsf{P}_{\sigma\theta}(\cdot)$

Denote for $\xi \in \mathcal{D}_{h^*}$:

$$\widehat{\partial}^2 \psi(\xi) := \partial^2 h^*(\xi) + \sigma \mathcal{A} \widehat{\partial}\, \mathsf{P}_{\sigma\theta}(x^k - \sigma \mathcal{A}^* \xi) \mathcal{A}^*$$

$\partial^2 h^*(\xi)$ = Clarke subdifferential of ∇h^* at ξ

$\widehat{\partial}\, \mathsf{P}_{\sigma\theta}(x^k - \sigma \mathcal{A}^* \xi)$ = "Jacobian" of $\mathsf{P}_{\sigma\theta}(\cdot)$

Define

$$V^0 := H^0 + \sigma \mathcal{A} U^0 \mathcal{A}^*$$

with $H^0 \in \partial^2 h^*(\xi)$, $U^0 \in \widehat{\partial}\, \mathsf{P}_{\sigma\theta}(x^k - \sigma \mathcal{A}^* \xi)$.

$V^0 \in \widehat{\partial}^2 \psi(\xi)$ and $V^0 \succ 0$

Semismooth Newton method.

Given $\bar{\eta} \in (0,1)$, $\tau \in (0,1]$, and $\delta \in (0,1)$. Choose $\xi^0 \in \mathcal{D}_{h^*}$. Iterate

1. Choose $V_j \in \widehat{\partial}^2 \psi(\xi^j)$. Find an approx. sol. $d^j \in \mathcal{Y}$ to

$$V_j(d) = -\nabla\psi(\xi^j)$$

$$\text{s.t.} \qquad \|V_j(d^j) + \nabla\psi(\xi^j)\| \leq \min(\bar{\eta}, \|\nabla\psi(\xi^j)\|^{1+\tau}).$$

2. (Line search) Fix $\varsigma \in (0, 1/2)$. Find the first nonnegative integer m s.t. $\xi^j + \delta^m d^j \in \mathcal{D}_{h^*}$ and

$$\psi(\xi^j + \delta^m d^j) \leq \psi(\xi^j) + \varsigma\delta^m \langle \nabla\psi(\xi^j),\, d^j \rangle$$

3. Set $\xi^{j+1} = \xi^j + \delta^m\, d^j$.

Theorem 29

Assume that $\nabla h^*(\cdot)$ *and* $\mathrm{Prox}_{\sigma p}(\cdot)$ *are strongly semismooth on* $\mathcal{D}_{h^*}$ *and* $\mathcal{X}$. *Then* $\{\xi^j\}$ *converges to the unique optimal solution* $\bar{\xi} \in \mathcal{D}_{h^*}$ *and*

$$\|\xi^{j+1} - \bar{\xi}\| = O(\|\xi^j - \bar{\xi}\|^{1+\tau}).$$

Implementable stopping criteria: the stopping criteria (A) and (B) $\Rightarrow$

$$(\mathrm{A}') \quad \|\nabla \psi_k(\xi^{k+1})\| \leq \frac{\varepsilon_k}{\max(1, \sqrt{\sigma_k/\alpha_h})}$$

$$(\mathrm{B}') \quad \|\nabla \psi_k(\xi^{k+1})\| \leq \frac{\delta_k}{\max(1, \sigma_k/\alpha_h)} \min\{1, \sigma_k \|\mathcal{A}^* \xi^{k+1} + u^{k+1}\|\}$$

An APG for (P).

Given $x^0 \in \operatorname{dom} p$, let $y^1 = x^0$ and $t_1 = 1$. Iterate

1. Compute

$$x^k = \arg\min_x \left\{ \begin{array}{l} \theta(x) + h(\mathcal{A}y^k) + \langle \mathcal{A}^* \nabla h(\mathcal{A}y^k),\, x - y^k \rangle \\ + \dfrac{L}{2}\|x - y^k\|^2 \end{array} \right\}.$$

2. Compute $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$

$$y^{k+1} = x^k + \frac{t_k - 1}{t_{k+1}}(x^k - x^{k-1}).$$

Recall h has a $1/\alpha_h$-Lipschitz continuous gradient:

$$L = \lambda_{\max}(\mathcal{A}^*\mathcal{A})/\alpha_h$$

An (linearized) ADMM for (D).

Given $(\xi^0, u^0, x^0) \in \mathcal{D}_{h^*} \times \mathrm{dom}\, p^* \times \mathcal{X}$, $\sigma > 0$, $\tau \in (0, (1+\sqrt{5}\,)/2)$.
Iterate

1. Compute

$$\begin{aligned} \xi^{k+1} &= \arg\min \left\{ \mathcal{L}_\sigma(\xi, u^k; x^k) + \frac{\sigma}{2}\|\xi - \xi^k\|^2_{\mathcal{T}} \right\} \\ u^{k+1} &= \arg\min\{\mathcal{L}_\sigma(\xi^{k+1}, u; x^k)\}. \end{aligned}$$

2. Compute $x^{k+1} = x^k - \tau\sigma(\mathcal{A}^*\xi^{k+1} + u^{k+1})$.

Classical ADMM: $\mathcal{T} = 0$

Linearized ADMM: $\mathcal{T} = \lambda_{\max}(\mathcal{A}\mathcal{A}^*)\mathcal{I} - \mathcal{A}\mathcal{A}^*$

Specialized matrix free interior point method (mfIPM) [Fountoulakis et al. 2014] for lasso
$x^* = \arg\min \frac{1}{2}\|\mathcal{A}x - b\|^2 + \lambda_1 \|x\|_1$, reformulated as:

$$z^* = \arg\min \left\{ \lambda_1 \langle \mathbf{1}_{2n},\, z \rangle + \frac{1}{2}\|\mathcal{F}^* z - b\|^2 \;\middle|\; z \geq 0 \right\}$$

$$z^* = [u^*; v^*], \quad \mathcal{F}^* = [\mathcal{A},\, -\mathcal{A}], \quad x^* = u^* - v^*$$

Linear system to solve:

$$\begin{pmatrix} \mathcal{F}\mathcal{F}^* & -\mathcal{I}_{2n} \\ S & Z \end{pmatrix} \begin{pmatrix} dS \\ dZ \end{pmatrix} = \begin{pmatrix} \mathsf{rhs}_1 \\ \mathsf{rhs}_2 \end{pmatrix}$$

Using Schur complement formula to get $2n \times 2n$ system

$$(Z^{-1}S + \mathcal{F}\mathcal{F}^*)dZ = \mathsf{rhs}_1 + Z^{-1}\mathsf{rhs}_2$$

mfIPM: solve the above system by PCG with specialized preconditioner for lasso problems!

Newton system for LASSO

LASSO: $\min\left\{\frac{1}{2}\|\mathcal{A}x - b\|^2 + \lambda_1\|x\|_1\right\}$

$h(y) = \frac{1}{2}\|y - b\|^2$ and $\theta(x) = \lambda_1\|x\|_1$

Proximal mapping of θ: easy to compute (soft thresholding)

Newton System:

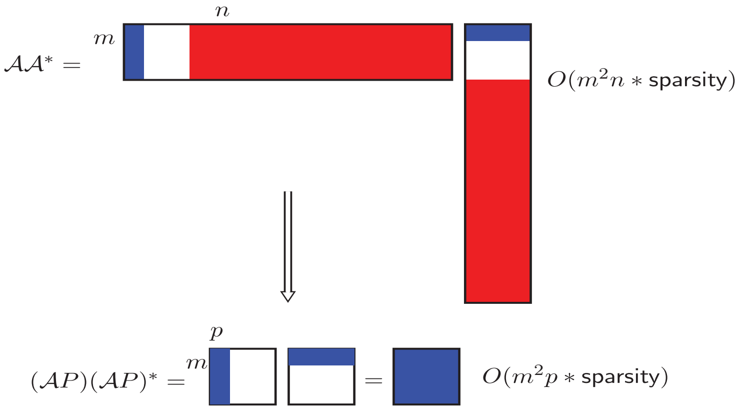
$$(I_m + \sigma\mathcal{A}P\mathcal{A}^*)d\xi = \text{rhs}$$

$P \in \widehat{\partial}\mathrm{P}_{\sigma\theta}(x^k - \sigma\mathcal{A}^*\xi)$: a diagonal matrix with $0, 1$ entries!

$$P_{ii} = \begin{cases} 0 & \text{if } |(x^k - \mathcal{A}^*\xi)_i| < \sigma\lambda_1 \\ 1 & \text{otherwise} \end{cases}$$

Using sparsity to solve the system efficiently!

Exploit the second order sparsity:



If $p \ll m$, use Sherman-Morrison-Woodbury formula to invert:

$$(I_m + UU^*)^{-1} = I_m - U(I_p + U^*U)^{-1}U^*$$

Need only to invert smaller $p \times p$ matrix: $(I_p + \sigma(\mathcal{A}P)^*(\mathcal{A}P))$

SSNAL for fused lasso and efficient implementation

Fused lasso: $\min\{\frac{1}{2}\|Ax-b\|^2 + \lambda_1\|x\|_1 + \lambda_2\|\mathcal{B}x\|_1\}$

$$\mathcal{B} = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix}$$

$h(y) = \frac{1}{2}\|y-b\|^2$ and $\theta(x) = \lambda_1\|x\|_1 + \lambda_2\|\mathcal{B}x\|_1$

For details: attend the talk of Xudong Li on Thursday

KKT residual:

$$\eta_{\mathrm{KKT}} := \frac{\|\tilde{x} - \mathsf{P}_\theta(\tilde{x} - (\mathcal{A}\tilde{x} - b))\|}{1 + \|\tilde{x}\| + \|\mathcal{A}\tilde{x} - b\|} \leq 10^{-6}.$$

Compare SSNAL with state-of-the-art solvers: mfIPM and FPC_AS

$(\mathcal{A}, b)$ taken from 11 Sparco collection [Van Den Berg, 2009]

$\lambda = \lambda_c \|\mathcal{A}^* b\|_\infty$ with $\lambda_c = 10^{-3}$ and 10^{-4}

Add 60dB noise to b in MATLAB: `b = awgn(b,60,'measured')`

max. iteration number: 20000 for FPC_AS

max. computation time: 7 hours

Numerical results for LASSO arising from compressed sensing

(a) our SsNAL (b) mfIPM
(c) FPC_AS: 1st order method based on FB splitting

$\lambda_c = 10^{-3}$		η_{KKT}	time (hh:mm:ss)
probname	$m;n$	a | b | c	a | b | c
p3poly	600;2048	2.3-7 | 9.5-7 | 9.2-7	48 | 1:09 | 29
blknheavi	1024;1024	6.3-7 | 9.2-7 | 1.3-1	01 | 01 | 55
srcsep1	29166;57344	9.5-7 | 7.3-7 | 9.7-7	7:20 | 42:34 | 13:25
soccer1	3200;4096	2.6-9 | 6.3-7 | 5.2-1	02 | 03 | 13:51
blurrycam	65536;65536	8.2-7 | 6.5-7 | 3.6-8	02 | 09 | 03
blurspike	16384;16384	2.3-7 | 9.5-7 | 7.4-4	05 | 05 | 6:38
$\lambda_c = 10^{-4}$			
p3poly	600;2048	2.5-7 | 8.3-7 | 1.1-2	2:04 | 2:13 | 42:52
blknheavi	1024;1024	1.9-7 | 8.7-7 | 4.6-3	01 | 01 | 49
srcsep1	29166;57344	3.0-7 | 9.5-7 | 9.9-7	17:23 | 3:31:08 | 32:28
soccer1	3200;4096	5.1-8 | 4.3-7 | 5.2-1	02 | 02 | 13:23
blurrycam	65536;65536	7.9-7 | 9.7-7 | 1.3-7	05 | 1:35 | 08
blurspike	16384;16384	3.4-7 | 7.4-7 | 8.3-5	11 | 08 | 6:43

Numerical results for LASSO arising from sparse regression

11 large scale instances $(\mathcal{A}, b)$ from LIBSVM [Chang and Lin, 2011]

For the data **pyrim, triazines**, ... ,**housing, mpg**, we expand their original features by using polynomial basis functions [Huang et al., 2010]. The last digit in **pyrim5** means order 5 polynomial is used

$\lambda_c = 10^{-3}$		$\eta_{\rm KKT}$	time (hh:mm:ss)
probname	$m; n$	a | b | c	a | b | c
E2006.train	16087; 150360	1.7-11 | 3.9-7 | 9.0-4	01 | 11 | 1:34:40
log1p.E2006.train	16087; 4272227	1.8-7 | 4.7-8 | 4.7-1	26 | 41:01 | 7:00:20
E2006.test	3308; 150358	6.0-7 | 2.9-7 | 3.7-4	00 | 05 | 33:30
log1p.E2006.test	3308; 4272226	5.6-7 | 1.5-8 | 9.8-1	22 | 37:24 | 7:00:01
pyrim5	74; 201376	5.3-7 | 3.2-7 | 8.8-1	05 | 21:27 | 2:01:23
triazines4	186; 635376	3.3-7 | 8.8-1 | 9.1-1	28 | 53:30 | 7:30:17
abalone7	4177; 6435	8.7-7 | 3.5-7 | **Error**	02 | 49 | **Error**
housing7	506; 77520	2.8-7 | 8.1-1 | 8.4-1	04 | 5:13:19 | 1:41:01
mpg7	392; 3432	6.3-9 | 6.5-7 | **Error**	00 | 04 | **Error**

2nd order method is important for designing robust solvers!

(a) our SSNAL
(b) state-of-the-art: APG based solver, SLEP [Liu et al 2009]
(c1) ADMM (classical) (c2) ADMM (linearized)

Parameters: $\lambda_1 = \lambda_c \|\mathcal{A}^* y\|_\infty,\ \lambda_2 = 2\lambda_1,$ tol $= 10^{-4}$

Problem: triazines 4, $m = 186$, $n = 635376$

	iter	time (hh:mm:ss)
λ_c | nnz | η_C	a | b | c1 | c2	a | b | c1 | c2
10^{-1} ; 164; 2.4-2	10 | 6448 | 3461 | 8637	18 | 26:44 | 28:42 | 46:35
10^{-2} ; 1004; 1.7-2	13 | 11820 | 3841 | 19596	22 | 48:51 | 24:41 | 1:22:11
10^{-3} ; 1509; 1.2-3	16 | 20000 | 4532 | 20000	31 | 1:16:11 | 38:23 | 1:29:48
10^{-4} ; 2205; 2.6-4	21 | 20000 | 6353 | 20000	48 | 1:14:57 | 53:58 | 1:27:11
10^{-5} ; 2420; 6.4-5	24 | 20000 | 14384 | 20000	1:01 | 1:26:39 | 1:49:44 | 1:35:36

SSNAL is vastly superior to first-order methods: APG, ADMM (classical), ADMM (linearized)

ADMM (linearized) needs many more iterations than ADMM (classical)

As problems get more complex, 2nd order method is even more important!

(SDP) $\quad \min \Big\{ \langle C,\, X \rangle \mid \mathcal{A}X - b = 0, X \in \mathbb{S}^n_+ \Big\}$

(SDD) $\quad \max \Big\{ -p^*(-S) + \langle b,\, y \rangle \mid \mathcal{A}^* y + S - C = 0, \;\; \xi \in \mathbb{R}^m \Big\}$

where $p(X) = \delta_{\mathbb{S}^n_+}(X)$. Let X^k be the multiplier associated with the dual equality at the kth iteration. Let $G^k = C - \sigma^{-1} X^k$. The augmented Lagrangian subproblem for (SDD) is

$$
\begin{aligned}
& \min_{S,y} \Big\{ p^*(-S) + \frac{\sigma}{2} \| S + \mathcal{A}^* y - G^k \|^2 - \langle b,\, y \rangle \Big\} \\
= & \min_{y} \Big\{ -\langle b,\, y \rangle + \min_{S} \big\{ p^*(-S) + \frac{\sigma}{2} \| S + \mathcal{A}^* y - G^k \|^2 \big\} \Big\} \\
= & \min_{y} \Big\{ \Phi^k(y) := -\langle b,\, y \rangle + \sigma \mathsf{M}_{\sigma^{-1} p^*}(\mathcal{A}^* y - G^k) \Big\}
\end{aligned}
$$

$$
\begin{aligned}
\nabla \Phi^k(y) &= -b + \sigma \mathcal{A} \nabla \mathsf{M}_{\sigma^{-1} p^*}(\mathcal{A}^* y - G^k) = -b + \mathcal{A} \mathsf{P}_{\sigma p}(\sigma \mathcal{A}^* y - \sigma G^k) \\
&= -b + \mathcal{A}\, \Pi_{\mathbb{S}^n_+}(\sigma \mathcal{A}^* y - \sigma G^k)
\end{aligned}
$$

A semismooth Newton-CG method for the subproblem

Solve $\nabla\Phi^k(y) = b - \mathcal{A}\Pi_{\mathbb{S}^n_+}(U) = 0, \quad U = \sigma\mathcal{A}^* y - \sigma G^k$.

$\nabla\Phi^k(y)$ is not differentiable, but is strongly semismooth [Sun-Sun]. At the current iteration, y_l, we solve a generalized Newton equation:

$$\mathcal{H}\Delta y \approx \nabla\Phi^k(y_l), \quad \text{where } \mathcal{H}\Delta y = \sigma\mathcal{A}\Pi'_{\mathbb{S}^n_+}(U)[\mathcal{A}^*\Delta y] \tag{22}$$

From eigenvalue decomp: $U = QDQ^T$ with $d_1 \geq \cdots \geq d_r \geq 0 > d_{r+1} \geq \cdots \geq d_n$, we choose

$$\Pi'_{\mathbb{S}^n_+}(U)[M] = Q(\Omega \circ (Q^T M Q))Q^T, \tag{23}$$

where $\Omega_{ij} = (d_i^+ - d_j^+)/(d_i - d_j)$. For $\gamma = \{1, \ldots, r\}$ and $\bar{\gamma} = \{r+1, \ldots, n\}$, we have

$$\Omega = \begin{bmatrix} E_{\gamma\gamma} & \Omega_{\gamma\bar{\gamma}} \\ \Omega_{\bar{\gamma}\gamma} & 0 \end{bmatrix}.$$

The structure in Ω allows for efficient computation of rhs of (23), and hence matrix-vector multiply for CG in solving (22)

(SDP+) $\min\left\{\langle C, X\rangle \mid \mathcal{A}(X) = b,\ l \le \mathcal{B}(X) \le u,\ X \in \mathbb{S}^n_+,\ X \in \mathcal{P}\right\}$

where $\mathcal{P} = \{X \in \mathbb{S}^n \mid L \le X \le U\}$ imposes bounds on X.

$$\begin{array}{ll} \text{(D)} & \min \delta^*_{\mathcal{P}}(-Z) + \delta^*_{\mathcal{Q}}(-v) + \langle -b,\, y\rangle \\ \text{s.t.} & \mathcal{A}^*(y) + \mathcal{B}^*(v) + S + Z = C,\ S \in \mathbb{S}^n_+ \end{array}$$

where $Q = \{s \in \mathbb{R}^p \mid l \le s \le u\}$.
While the problem (SDP+) has only a single block X, our solver can solve the following more general problem with N blocks of variables:

$$\begin{array}{ll} \min & \sum_{j=1}^N \langle C^{(j)},\, X^{(j)}\rangle \\ \text{s.t.} & \sum_{j=1}^N \mathcal{A}^{(j)}(X^{(j)}) = b, \quad l \le \sum_{j=1}^N \mathcal{B}^{(j)}(X^{(j)}) \le u \\ & X^{(j)} \in \mathcal{K}^{(j)},\ X^{(j)} \in \mathcal{P}^{(j)},\ j = 1, \ldots, N \end{array}$$

Robustness of SDPNAL+

For the class where $\mathcal{P} = \{X \geq 0\}$, $\mathcal{B} = 0$ (no linear inequalities)

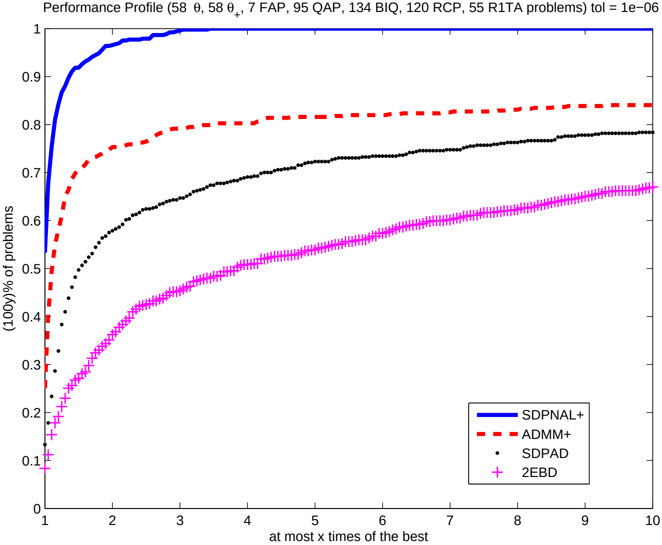
$$\eta_{\mathsf{KKT}} := \max \left\{ \begin{array}{c} R_P, R_D, R_{\mathbb{S}^n_+}(X), R_{\mathcal{P}}(X), R_{\mathbb{S}^n_+}(S), R_{\mathcal{P}}(Z), \\ R(\langle X,\, S\rangle), R(\langle X,\, Z\rangle) \end{array} \right\} \leq 10^{-6}.$$

We compare the performance of our SDPNAL+ and ADMM+ with the direct ADMM implemented in SDPAD [Wen et al.] and 2EBD-HPE [Monteiro et al.]

Numbers of problems which are solved to the accuracy $\eta_{\mathsf{KKT}} \leq 10^{-6}$

problem set (No.)	SDPNAL+	ADMM+	SDPAD	2EBD
θ (58)	58	56	53	53
θ_+ (58)	58	58	58	56
FAP (7)	7	7	7	7
QAP (95)	95	39	30	16
BIQ (134)	134	134	134	134
RCP (120)	120	120	114	109
R1TA (55)	55	42	47	18
Total (527)	527	456	443	393

Performance profiles of SDPNAL+, ADMM+, SDPAD and 2EBD



Implemented the algorithms in MATLAB.
Runs perform on a 6 cores Linux Server with 12 Intel Xeon processors at 2.67 GHz and 32G RAM.

Stop SDPAD and 2EBD after 25000 iterations or 99 hours.

Prob	$m; n$	η SDPAD|2EDB|SDPNAL+	time (hour:minute)
1dc.2048	58368; 2048	9.9-7| 9.9-7| 9.9-7	14:00| 16:04| 5:50
fap36	4110+$\mathcal{N}$; 4110	9.9-7| 9.9-7| 9.5-7	78:43| 43:37| 23:07
nug30	1393+$\mathcal{N}$; 900	1.1-5| 1.7-5| 9.6-7	4:58| 5:39| 0:45
tai30a	1393+$\mathcal{N}$; 900	4.6-6| 1.3-5| 9.9-7	6:09| 6:00| 0:29
nonsym(6,5)	194480; 1296	9.9-7| **1.6-3**| 5.2-7	2:59| 11:24| 0:05
nsym_rd[40,40,40]	672399; 1600	**3.7-4**| **5.1-4**| 8.6-7	13:56| 22:41| 0:14
nonsym(12,4)	12.32M; 9261	**9.8-3**| **5.2-3**| 5.7-8	99:00| 99:00| 14:22

Results show that it is essential to use second-order information to solve hard problems!

Summary for SDPNAL+

- We have tested SDPNAL+ on about 520 SDPs from θ, θ_+, QAP, binary QP, rank-1 tensor approximation, etc
- When the problems are primal-dual nondegenerate, SDPNAL+ can efficiently solve large SDPs to high accuracy. SDPAD and 2EDB also performed well, though SDPNAL+ is often more efficient.
- Many of the tested SDPs are degenerate, but SDPNAL+ can still solve them accurately with $\eta < 10^{-6}$. Other hand, SDPAD and 2EDB were not able to solve many such problems.

Thank you for your attention!